UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

DETECCIÓN DE OPOSICIÓN SEMÁNTICA MEDIANTE PATRONES SINTÁCTICOS YUXTAPUESTOS

TESIS
QUE PARA OPTAR POR EL GRADO DE
DOCTOR EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

PRESENTA:
ALEJANDRO ESTEBAN PIMENTEL ALARCÓN

TUTOR
DR. GERARDO EUGENIO SIERRA MARTÍNEZ INSTITUTO DE INGENIERÍA

MIEMBROS DEL COMITÉ TUTOR
DR. GIBRAN FUENTES PINEDA IIMAS
DR. ALEXANDER GELBUKH IPN

CIUDAD DE MÉXICO, ENERO 2022

Universidad Nacional Autónoma de México

Posgrado en Ciencia e Ingeniería de la Computación

Inteligenica Artificial

Detección de oposición semántica mediante patrones sintácticos yuxtapuestos

T E S I S

que para obtener el grado de:

Doctor en Ciencia e Ingeniería de la Computación

presenta:

Alejandro Esteban Pimentel Alarcón

director de tesis:

Dr. Gerardo Eugenio Sierra Martínez

comité tutor:

Dr. Gibran Fuentes Pineda

Dr. Alexander Gelbukh

UNAM
POSGRADO
Ciencia e Ingeniería de la Computación

Ciudad de México Enero, 2022

Detección de oposición semántica mediante patrones sintácticos yuxtapuestos

por

Alejandro Esteban Pimentel Alarcón

Mtro. en Ciencia e Ingeniería de la Computación (2017)
Ing. Mecatrónico, Universidad Nacional Autónoma de México (2015)

Tesis presentada para obtener el grado de

Doctor en Ciencia e Ingeniería de la Computación

en el

Posgrado en Ciencia e Ingeniería de la Computación

Universidad Nacional Autónoma de México

Ciudad de México. Enero, 2022

Investigación es lo que se hace
cuando no se sabe lo que se hace

Wernher von Braun

Agradecimientos

Este trabajo fue posible gracias al apoyo del Consejo Nacional de Ciencias y Tecnología (CONACYT) por la beca de manutención 483759 que me permitió realizar los estudios de doctorado. (Becario 334720)

Este trabajo recibió también el apoyo del proyecto UNAM-PAPIIT "Detección automática de pensamientos suicidas en textos" con clave AG400119.

Índice general

Índice de figuras

Detección de oposición semántica mediante patrones sintácticos yuxtapuestos

por

Alejandro Esteban Pimentel Alarcón

Resumen

Junto con el aumento de la producción y la velocidad de generación de información textual, crece también el interés y la necesidad de ser capaces de manejarla automáticamente. Los llamados *word embeddings* han sido una herramienta fundamental para la meta de procesar de forma automática el lenguaje que se ha podido aplicar en un gran número de idiomas. Sin embargo, presentan ciertas limitantes, principalmente cuando se trata de distinguir entre diferentes formas de relación entre términos asociados. Para lidiar con esta limitante es común ver en la literatura el uso de ontologías y otros recursos léxicos, pero dichos recursos son difíciles, tardados y costosos de producir y mantener.

La presente investigación busca aminorar la carga de la obtención de relaciones léxicas, particularmente las relaciones de oposición, que son las más problemáticas para los *word embeddings*. Para lograr lo anterior, se presenta el desarrollo de una metodología basada en el uso de patrones sintácticos yuxtapuestos que aprovecha la gran disponibilidad de texto para encontrar relaciones propias de una base léxica. De esta manera, se logra tanto enriquecer dichas bases como un sistema de *embeddings* sin recurrir a recursos lingüísticos externos, lo que muestra cierto grado de independencia del idioma, es decir, que la metodología puede funcionar en idiomas que se pueden tokenizar y usen pre o post-posiciones.

La aplicación de la metodología también mostró su capacidad de extraer pares de palabras con oposición (antónimos y co-hipónimos) según una percepción humana, ya que dichos pares son, en su mayoría, coincidentes con el etiquetado que se llevó a cabo. Además, también demostraron ser una fuente adicional capaz de enriquecer representaciones vectoriales en un ámbito que resulta complicado de adquirir para sistemas distribucionales. Las aportaciones de nuestro sistema fueron aplicadas a GloVe mediante *counter-fitting* y comparadas con otras metodologías; se hizo una evaluación con pares obtenidos mediante un sistema basado en patrones simétricos (el método más cercano al que aquí se propone) y con pares obtenidos de recursos léxicos (WordNet y PPDB). Se observaron, para los resultados del presente trabajo, una mejora de alrededor de 7 % de precisión. Por último, se evaluó también la aportación de los pares de oposición sobre la tarea de minería de opinión con negación; para esta tarea, los pares se explotaron en la búsqueda de una representación en forma de transformación lineal para la negación, en donde los opuestos fueron usados como conjunto de entrenamiento de dicha transformación, es decir, una palabra, negada, se espera que obtenga como resultado su opuesto en el espacio vectorial.

Detección de oposición semántica mediante patrones sintácticos yuxtapuestos

by

Alejandro Esteban Pimentel Alarcón

Abstract

As the production and speed of production of textual information increases, so does the interest and the demand to be able to handle it automatically. The so-called word embeddings have been a fundamental tool for the goal of automatic language processing and have been applied in a large number of languages. However, they present certain limitations, mainly when it comes to distinguishing between different forms of relationships between associated terms. To deal with this limitation it is common to see the use of ontologies and other lexical resources in the literature, but such resources are difficult, time-consuming and costly to produce and maintain.

The present research seeks to lessen the burden of obtaining lexical relations, particularly opposition relations, which are the most problematic for word embeddings. To achieve the mentioned objective, we present the development of a methodology based on the use of juxtaposed syntactic patterns that takes advantage of large text availability to find relations similar to those of a lexical base. In this way, it is possible to enrich such bases as well as a system of embeddings without resorting to external linguistic resources, thus showing a certain degree of language independence, i.e., the methodology can work in languages that can be tokenized and that use pre- or post-positions.

The application of the methodology also demonstrated its ability to extract pairs of opposite words (antonyms and cohyponyms) according to a human perception, since these pairs are, for the most part, coincident with the labeling performed. Moreover, they also proved to be a source of information capable of enriching vector representations in a scope difficult to acquire for distributional systems. The contributions of our system were applied to GloVe by counter-matching and compared with other methodologies: an evaluation was performed with pairs obtained using a system based on symmetric patterns (the closest method to the one proposed here) and with pairs obtained from lexical resources (WordNet and PPDB). For the results of the present work, an accuracy improvement of about 7 % was observed. Finally, the contribution of opposition pairs on the opinion mining task with negation was also evaluated; for this task, the pairs were exploited in the search for a negation representation in the form of a linear transformation, where the opposites were used as training set of such transformation, i.e., a word, negated, is expected to obtain as a result its opposite in the vector space.

Capítulo 1

Introducción

El ser humano usa la lengua como principal medio de comunicación, una herramienta con la que es capaz de transmitir sus ideas. Con la llegada de la escritura, el conocimiento pudo ser plasmado para tener trascendencia a través del tiempo. La información se conserva, se organiza, se replica y se extiende; en la actualidad existen grandes cantidades de datos textuales y, desde la llegada del internet, la generación de texto se ha incrementado a dimensiones sin precedentes. Un ejemplo claro de esto es el informe "A day of data" del *World Economic Forum*, en el que se resalta que se estima que habrá 44 Zettabytes de información a nivel mundial para el 2020; es decir, 40 veces más bytes que estrellas en el universo[1].

Son necesarios métodos para trabajar con toda la información, en su mayoría escrita, que se crea y se ha creado. Dicha tarea, imposible de llevar a cabo mediante recursos humanos, requiere la automatización del manejo de la lengua escrita, es por ello que se recurre al procesamiento del lenguaje natural (PLN), área de la inteligencia artificial (IA) que se ocupa de la formulación de modelos computacionales para la comunicación entre personas y máquinas.

El área de procesamiento de lenguaje natural ha visto un gran impulso en su desempeño desde la aparición de las representaciones distribucionales, que han permitido cierta codificación semántica en espacios vectoriales densos. Una de las principales fortalezas de estos sistemas es que no necesitan datos etiquetados para su obtención, les basta tan solo texto plano, del cual hay

[1] `https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f`

una gran disponibilidad. Estos sistemas, así como todos sus derivados, se basan en la hipótesis de distribución de Harris [1954], que en esencia plantea que: palabras similares ocurren en contextos similares.

Dichos sistemas distribucionales presentan propiedades muy interesantes que reflejan ciertas relaciones semánticas entre palabras. Como ejemplo se tiene un sistema Euclidiano en que cada palabra toma una posición dentro de un espacio vectorial, es decir, un punto, y la posición relativa entre los puntos es lo que logra la codificación. En este sentido, las palabras con mayor relación estarán mas cerca entre ellas, y no solo eso, sino que se ha encontrado que se pueden llevar a cabo analogías con diferencias de vectores, por ejemplo: $V_{rey} - V_{hombre} \approx V_{reina} - V_{mujer}$ [Mikolov *et al.*, 2013a]. Donde V_x representa el vector para la palabra x.

A pesar de esto, existen muchos tipos de similitud que los sistemas distribucionales no son capaces de distinguir y quedan agrupados en una similitud general, comúnmente representada como cercanía en el espacio vectorial. Ello es debido al núcleo de su proceso, que se basa en la relación de palabras mediante ventanas de contexto: dentro de mismos contextos pueden ser usadas palabras con otras relaciones. Un buen ejemplo son los antónimos, palabras que son similares en categoría, pero opuestas dentro de la misma. Los antónimos se verán en contextos similares al igual que los sinónimos, lo cual impide que un sistema distribucional los distinga.

Son principalmente los antónimos y las relaciones de oposición las que son de mayor interés para este trabajo, ya que son las palabras menos "similares" que son agrupadas por los sistemas distribucionales. Por lo general, para lidiar con este tipo de relaciones, se suele recurrir al uso de bases léxicas: bases de datos de naturaleza ontológica como WordNet, donde se recopilan y distinguen los diferentes tipos de relaciones que existen entre las palabras. Lamentablemente, este tipo de recursos son costosos de producir, y es bien sabido que su cobertura es bastante limitada. Por ello, en esta investigación nos damos a la tarea de encontrar un método para detectar y extraer oposición de forma automática.

1.1. Objetivos

El presente trabajo tiene como meta contribuir en aminorar la carga de la construcción de recursos léxicos, particularmente con respecto a relaciones de oposición. Para ello, se busca aprovechar la extensa disponibilidad de texto con la intención de encontrar pares de palabras relacionados que se puedan aprovechar para enriquecer tanto bases léxicas como sistemas de *embeddings*, razón por la que se está enfocando sobre las relaciones de oposición (co-hiponímia y antonimia) que tienden a ser las más elusivas para todos los sistemas distribucionales.

1.1.1. Objetivos específicos

Para este objetivo general se toman en cuenta los siguientes objetivos específicos.

- Diseñar un método capaz de extraer patrones automáticamente.
- Extraer grupos de palabras relacionadas obtenidas mediante la metodología diseñada.
- Evaluar la naturaleza de la información que se obtenga mediante dicha extracción.
- Enriquecer *word embeddings* mediante la adición de la información extraída gracias a los grupos de palabras formados.
- Evaluar la aportación.

1.2. Hipótesis

Para cumplir con estos objetivos, en este trabajo se parte de la premisa de que dentro de la lengua escrita se pueden encontrar estructuras léxicas indicadoras de oposición que se usan de forma constante para vincular pares de palabras; y que tanto las regularidades como la información de oposición que aportan se pueden aprovechar para impactar directamente de forma positiva en los sistemas de representación vectorial de palabras.

1.3. Estructura de la tesis

Para el logro de los objetivos y la contrastación de la hipótesis, la investigación constará de este y otros siete capítulos. En el segundo se desarrollarán a detalle las relaciones léxicas del lenguaje, dentro de las cuáles se hará principal énfasis en la relación que compete a esta investigación: la oposición, que como se ha mencionado anteriormente, es de gran interés debido a que las palabras opuestas conservan los mismos contextos entre sí. Ya que al mantener los mismos contextos son agrupados como similares en las representaciones distribucionales, en el tercer capítulo se abordará una revisión de dichas representaciones vectoriales de palabras basadas en distribuciones léxicas y se analizarán las formas en que han sido enriquecidos con relaciones léxicas externas. Los sistemas distribucionales, junto con muchas otras técnicas del PLN, se suelen ver beneficiados cuando, dentro de sus procesos, se filtran y no se toman en cuenta las palabras más comunes y funcionales (palabras como los artículos o las preposiciones); sin embargo, este tipo de palabras sí se usan cuando se buscan patrones o regularidades en el lenguaje, patrones que han dado lugar exitosamente a relaciones como la hiperonimia (como se puede observar en los ejemplos de las secciones 2.3 y 2.4.3); estas observaciones nos llevaron a desarrollar los experimentos preliminares que se muestran en el cuarto capítulo, una exploración sobre estructuras generales de palabras comunes que dieron lugar al hallazgo de fenómenos interesantes sobre los cuáles se desarrolló finalmente la metodología.

En el quinto capítulo se estudiará paso a paso la metodología que se desarrolló para la detección de oposición semántica mediante patrones sintácticos yuxtapuestos. Mientras que en el sexto capítulo se muestran experimentos y los resultados de cada uno de los módulos del proceso. En el séptimo capítulo se mostrarán las evaluaciones a las que se sometieron los resultados obtenidos, tanto para los co-hipónimos como los antónimos. La evaluación consiste en dos partes, una parte mediante un etiquetado humano para contrastar los pares de palabras con la percepción de oposicón y la segunda parte consiste en una evaluación automática mediante el enriquecimiento de un sistema de representación distribucional. Finalmente, se presentarán las conclusiones de la investigación.

Capítulo 2

Sobre la oposición y otras relaciones

El estudio del lenguaje natural presenta un punto de partida muy natural para ser segmentado y analizado de una forma modular: las palabras; o más formalmente: unidades léxicas. Se puede entender por unidad léxica aquellos compuestos que combinan forma y significado de una forma relativamente estable y discreta [Cruse *et al.*, 1986].

En general, las palabras pueden pertenecer a una de dos grandes categorías: léxicas o funcionales [Conde, 2005].

Palabras léxicas: También conocidas como palabras de contenido, son aquellas que tienen un significado completo, son la base del contenido semántico de las oraciones. Estas palabras están conformadas regularmente por conjuntos abiertos, es decir, son palabras que se crean constantemente, por lo que no es posible enumerarlas de forma efectiva. Dentro de los principales conjuntos que conforman a las palabras léxicas, se pueden nombrar, por ejemplo: Nombres, verbos, adjetivos y adverbios.

Palabras funcionales: Son aquellas que tienen propósitos gramaticales, es decir, funcionan principalmente como conectores para las palabras léxicas. Estas palabras están conformadas regularmente por conjuntos cerrados, es decir, son palabras fijas que casi no observan cambios a lo largo del tiempo, por lo que en muchos casos se puede llegar a contar con listados finitos para identificarlas. Dentro de los principales conjuntos que conforman a

las palabras funcionales se pueden nombrar, por ejemplo: Artículos, preposiciones, conjunciones, determinantes y pronombres.

Como se puede observar, dentro de cada una de las dos grandes categorías se pueden encontrar otro tipo de clasificación de palabras (verbos, adjetivos, artículos, etc.). Dichas clases corresponden a las partes de la oración (POS por sus siglas en inglés) que puede tomar una palabra; cada una de las partes de la oración corresponde a conjuntos de palabras que presentan propiedades gramaticales similares y un comportamiento sintáctico parecido. La información de POS de las palabras puede aportar una dimensión extra para el análisis de estructuras y, como se verá con más detalle en los siguientes capítulos, es una herramienta de utilidad para muchas tareas de PLN.

La segmentación en unidades léxicas permite estudiar, por un lado, el significado relativamente aislado del significado de las palabras (como podríamos observar en un diccionario), y por otro lado, la forma en que interactúan y se relacionan las palabras en los procesos de composición que logra la representación, y comunicación de estructuras de significado mas complejas. Como se verá en las siguientes secciones, principalmente para la detección de hiperonimia y de oposición entre palabras léxicas, ha demostrado ser efectivo el uso de patrones de palabras funcionales.

2.1. Principales relaciones léxicas

Como mencionan Cruse *et al.* [1986]: El significado de las unidades léxicas consiste en un número indefinido de relaciones con el contexto que al mismo tiempo constituye un solo ente unificado, por lo que se puede hablar de unidades léxicas que toman una relación semántica particular cuando se ponen en relación con otras unidades léxicas; el significado de cada unidad se revela entonces mediante las relaciones que guarda con su contexto.

En general, el estudio de relaciones léxicas está basado en los principios de pensamiento humano, todas las teorías lingüísticas usan la semántica como punto de partida [Conde, 2005]. En contraparte, tal como la semántica se usa como punto de partida, la estructura es utilizada

para obtener información sobre la semántica, de esta manera se pueden establecer dos tipos de conexiones entre palabras: verticales y horizontales [Orešković, 2019]. Las relaciones verticales se refieren a las conexiones que existen entre los significados de las palabras, mientras que las conexiones horizontales se refieren a las conexiones que existen entre la estructura y posición de las palabras.

Como veremos en las siguientes secciones, la búsqueda y extracción de relaciones léxicas se suele lograr mediante el aprovechamiento de relaciones horizontales, es decir de relaciones sintácticas. La sintaxis es la encargada de estudiar las reglas que se deben seguir para conformar mayores estructuras con significado y las palabras forman las unidades básicas para conformar dichas estructuras. En ocasiones, algunas estructuras sintácticas son muy comunes y pueden ser utilizadas como indicadoras de relaciones léxicas en particular, el ejemplo más común de esto es la estructura "X es un Y" que denota hiperonimia de Y a X (más detalles en la sección 2.3). Cuando hablamos de patrones sintácticos nos referimos a estas estructuras léxicas que se repiten frecuentemente y que pueden ser aprovechadas para la detección de relaciones entre las palabras variables que ocupan un lugar dentro del patrón (en el ejemplo, las palabras variables serían X y Y).

Cruse *et al.* [1986] afirman que, dentro de la amplitud de relaciones que se pueden formar entre las palabras, un conjunto acotado de relaciones se ha establecido en una posición central de discusión léxico-semántica (antónimos, hipónimos, sinónimos, entre otros), relaciones que en general también tienen congruencia espacial que se puede expresar en forma de conjuntos. En la figura 2-1 se muestran de forma gráfica los cuatro tipos de relaciones que se pueden presentar en un par de conjuntos: identidad, inclusión, superposición y disyunción. Mismas cuatro relaciones que pueden encontrar una equivalencia al momento de relacionar unidades léxicas, como se verá a continuación [Cruse *et al.*, 1986].

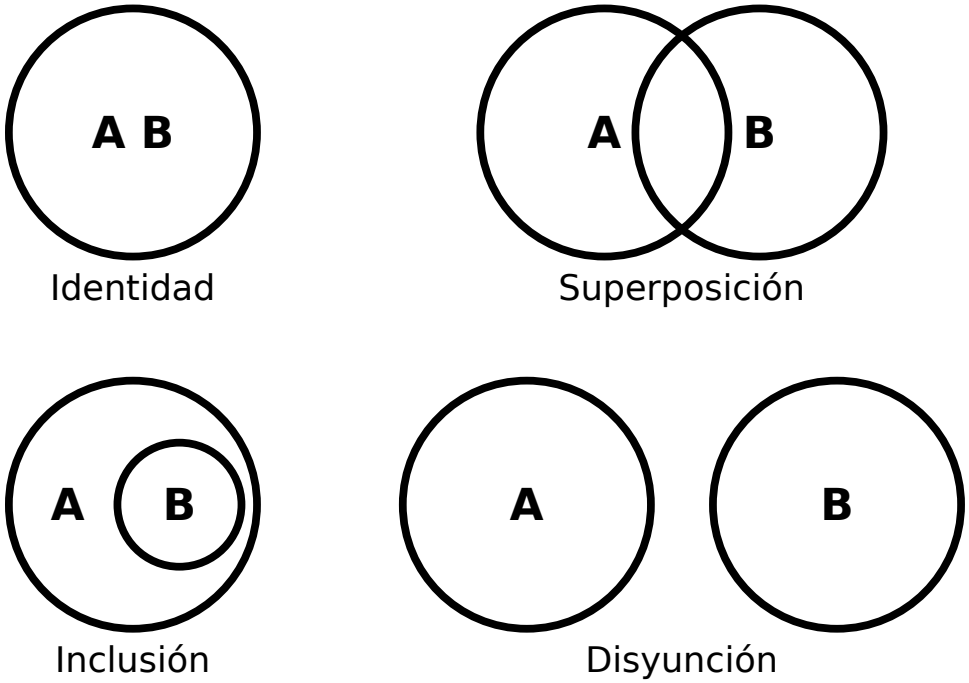

Figura 2-1: Diagramas de Venn ejemplificando los cuatro tipos de relaciones que se pueden presentar en un par de conjuntos.

2.2. Sinonimia

La sinonimia es la relación léxica que coincide con la relación de identidad (véase figura 2-1). Se puede considerar a un par de palabras X y Y como sinónimos siempre y cuando se cumpla que, en ciertos contextos, el significado de una oración S que contenga la unidad X no cambiará si se cambia el elemento X por el elemento Y. Es decir, son palabras que son típicamente intercambiables en algunos contextos, incluso si no lo son en otros [Yule, 2020]. La sinonimia perfecta, pares que se puedan intercambiar en cualquier contexto, es prácticamente inexistente [Edmonds y Hirst, 2002], característica que se puede extrapolar a las relaciones de antonimia.

La sinonimia es una relación léxica sumamente importante para el PLN y es probablemente la relación más estudiada y trabajada en la literatura. La extracción de sinónimos en una tarea de particular interés ya que puede ser utilizada en la mejora de tareas como la desambiguación de sentido, la construcción automática de ontologías y la substitución léxica, la cual a su vez

es fundamental para tareas como la recuperación de información, respuesta automática de preguntas o el resumen automático. La fuente principal de sinónimos en aplicaciones automáticas son las ontologías y tesauros [Roget, 1911; Fellbaum, 1995; Miller, 1998]. Pero este tipo de fuentes no logra buena cobertura ni generalización de dominios, además de que son difíciles y costosas de construir. Es por eso que también se buscan métodos para la extracción automática de sinónimos.

Por un lado, una categoría importante de métodos de extracción es mediante patrones que, como su nombre lo indica, consisten en el uso de patrones de palabras específicas que se usan en situaciones específicas. Si se puede vincular un patrón común para una relación entre palabras, entonces se convierte en un medio eficaz para la extracción de dicha relación sin la necesidad de que el sistema "comprenda" la semántica del contexto; tanto así, que es el método más utilizado para este tipo de tareas, como se verá a lo largo de este capítulo.

Entre otros trabajos, se tiene el de Ismail *et al.* [2017], en donde se usan sinónimos semilla para la extracción de sinónimos que más adelante usan para la tarea de construcción automática de ontologías. Otro ejemplo interesante de uso de patrones se presenta con el trabajo de Wang *et al.* [2010], en donde se parte de WordNet para la identificación de palabras relacionadas (su interés particular es para verbos); gracias a las palabras previamente relacionadas con WordNet, pueden llevar a cabo una búsqueda de patrones, los cuales definen como secuencias de palabras y sus respectivas etiquetas de partes de la oración de entre 3 y 5 elementos. De forma similar, en la investigación de Wang y Hirst [2012] también se busca la conformación de patrones, pero esta vez sobre un dominio específico: los diccionarios. Wang y Hirst [2012] se dan cuenta que, dentro de las definiciones, muchas veces se pueden encontrar sinónimos de los términos definidos dentro de la misma definición, por lo que diseñan un algoritmo que aproveche la estructura de las definiciones en el proceso de extracción de sinónimos.

Sin embargo, la naturaleza de los sinónimos de ser intercambiables entre ellos permite, para esta relación en particular, otra rama muy importante de algoritmos para su extracción: los algoritmos basados en sistemas distribucionales (más de este tema en la sección 3.2). En esencia, los métodos distribucionales hacen uso justamente de los contextos de las palabras para

buscar similitudes y relaciones semánticas que, en lugar de ser expresadas de forma explícita, se condensan en un espacio vectorial en el que se espera que palabras parecidas terminen en vecindades cercanas unas de otras. Estos métodos han marcado un antes y un después en el PLN, debido al gran éxito que han tenido y a las mejoras que han logrado, donde la extracción de sinónimos no es la excepción.

Zhang *et al.* [2017] aprovechan el algoritmo *word2vec* [Mikolov *et al.*, 2013a] para obtener un espacio vectorial como el antes mencionado a partir de la Wikipeda en inglés. Con ayuda de dicho espacio, buscan sinónimos mediante la proximidad de los vectores y el uso de algoritmos de *Spectral Clustering*[1] [Shi y Malik, 2000] para el agrupamiento de palabras en *clusters* de sinónimos.

Además, las técnicas que se basan en sistemas distribucionales abren las puertas a nuevas posibilidades, como la que presentan Hazem y Daille [2018]. En su investigación, atacan un problema poco abordado: la extracción de sinónimos multi-palabra. Cuando se trabaja con patrones, los términos multipalabra rompen los esquemas que se forman, pero los autores de esta investigación usan los vectores distribucionales y un conjunto semilla de sinónimos multipalabra para analizar cómo se relacionan los vectores dentro del espacio y extrapolarlo en búsqueda de nuevos sinónimos, independientemente de la cantidad de palabras que éstos requirieran.

Estos métodos también pueden ser utilizados dentro de dominios específicos, siempre y cuando se cuente con suficiente extensión para un entrenamiento aceptable del espacio vectorial. Henriksson *et al.* [2012] llevan a cabo precisamente esta estrategia para la búsqueda de sinónimos para el dominio médico, su método consiste en encontrar intersecciones en relaciones de varios procesos diferentes de vectorización para aumentar las probabilidades de encontrar sinónimos confiables.

No obstante, los métodos basados en sistemas distribucionales tienen una importante desventaja sobre los métodos basados en patrones. Como mencionan Nugumanova *et al.* [2019], uno de los mayores retos que se tienen en la extracción de sinónimos a partir de sistemas distribucionales, es poder diferenciarlos de las otras relaciones que también se presentan en la

[1] Algoritmos que agrupan puntos en espacios vectoriales usando eigenvectores o matrices derivadas de las distancias entre los puntos

extracción de palabras asociadas, pues los espacios vectoriales solo toman en cuenta una intercambiabilidad contextual, no semántica, lo que desemboca en que dentro del espacio se tienen no solo sinónimos, sino también antónimos y otras palabras solo asociadas (carro - gasolina) en las vecindades cercanas [Chen *et al.*, 2013]. En su investigación, Nugumanova *et al.* [2019] logran mejorar un poco la distinción de sinónimos en espacios vectoriales gracias a la combinación de diferentes métricas de similitud, pero los resultados aún así son relativamente pobres, rondando el 50 % de *F1-score*[2].

2.3. Hiperonimia

Esta es la relación léxica que corresponde a la relación de inclusión de los diagramas mostrados en la figura 2-1. En esta relación, un conjunto A incluye a un conjunto B. A diferencia del resto de las relaciones que se muestran en la figura, esta relación no es simétrica, y cada lado de esa asimetría recibe un nombre: A es hiperónimo de B, mientras que B es hipónimo de A; esta relación también está muy bien estudiada, y es la que obtiene mejores resultados cuando es trabajada con patrones léxicos.

Quizá el trabajo más conocido en el trabajo de extracción de hiperónimos sea el presentado por Hearst [1992]. En su artículo exponen patrones tal como "Y como un X" o "X y otros Y", que son indicativos de una relación de hiperonimia de Y hacia X. A partir de ese trabajo, los patrones se convirtieron en una herramienta estándar para la detección y extracción de relaciones entre palabras. El uso de patrones sintácticos permite extraer palabras relacionadas al buscar los términos que pueden tomar las posiciones variables dentro de los patrones: la estrategia entonces se convierte en el análisis de la lengua para la detección de buenos y, de preferencia, numerosos patrones. Como se verá a lo largo de esta sección, nuevas investigaciones han dado lugar a nuevos patrones y nuevas técnicas para encontrarlos; sin embargo, guardan mucha relación entre ellos, principalmente se puede notar que los patrones se componen, en su

[2]Métrica que se define como la media harmónica entre la precisión (P) y la cobertura (R) [Sasaki *et al.*, 2007]. Esto quiere decir que es una métrica que recompensa aquellos sistemas que cuando asignan una clase tienen una alta probabilidad de hacerlo correctamente (P) siempre y cuando no le hayan faltado muchas clases por detectar (R).

mayoría, de palabras funcionales.

El trabajo presentado por Snow *et al.* [2005] muestra una alternativa a la búsqueda manual de patrones para la extracción de hiperónimos. El método que se presenta en dicho trabajo consiste en el uso de aprendizaje máquina para reemplazar la etapa de análisis humano en el proceso de la construcción de patrones. Es decir, se toman pares de palabras previamente identificados como hiperónimos; dichos pares se toman como base para buscar oraciones en un corpus, tal que ambas palabras estén presentes; cada oración es automáticamente analizada con árboles de dependencias y, finalmente, se entrena un clasificador que usa el análisis como características.

Para facilitar la comparación, en la tabla 2-1 se muestran los resultados que obtiene Snow *et al.* [2005] con su método, dentro de los cuales se agrupan los patrones de Hearst en la parte superior, mientras que en la parte inferior de la misma tabla se muestran los patrones nuevos que fueron obtenidos por Snow *et al.* [2005].

Tabla 2-1: Patrones extraídos de forma automática por Snow *et al.* [2005]. En la parte superior se agrupan los patrones que coinciden con los encontrados por Hearst [1992].

X and other Y
X or other Y
Y such as X
Such Y as X
Y including X
Y, especially X
Y like X
Y called X
X is a Y
X a Y (appositive)

Como se puede ver, de los diez patrones que obtienen Snow *et al.* [2005] (y que se muestran en la tabla 2-1) seis coinciden con los ya mencionados patrones de Hearst, con la ventaja de

que estos patrones se obtienen automáticamente y, siempre que se cuente con pares semillas, se pueden aplicar para otros idiomas o contextos. Algunos contextos especializados cuentan con patrones particulares que pueden ser igualmente explotados, para el caso de los hiperónimos una fuente que ha demostrado ser muy productiva son las definiciones, ya que éstas están constituidas por un hiperónimo (o término cercano también conocido como *genus*) y una diferencia específica que separa al término definido de dicho hiperónimo y otros términos cercanos; en trabajos como el de Acosta *et al.* [2010] se aprovecha este vínculo que existe entre una definición y los hiperónimos para extraer pares hipónimo - hiperónimo semillas de forma automática.

El proceso inverso (hiperónimos - definiciones) también ha sido explotado con éxito. El uso de pares de palabras semillas para la exploración de patrones definitorios puede ser aprovechado sobre diferentes dominios con el fin de construir diccionarios de forma automática, tarea que es de gran interés, ya que los diccionarios también son costosos de construir y mantener, sobre todo los especializados [Dorantes *et al.*, 2017]. Podemos mencionar, como ejemplo, a Acosta *et al.* [2015], quienes presentan una metodología basada en patrones formados mediante el uso de hiperónimo - adjetivo (por ejemplo: enfermedad venérea) para la obtención de definiciones en el dominio médico. Sin embargo, la extracción de definiciones también puede ser aprovechada en contextos no especializados para ampliar o actualizar diccionarios generales; dichas aplicaciones ha motivado trabajos que aplican la extracción de patrones léxicos sobre la web [Dorantes *et al.*, 2017].

Sin embargo, podemos observar que para obtener este tipo de patrones es necesario satisfacer varias condiciones que requieren los métodos. Principalmente: pares de palabras previamente identificados como hiperónimos y un analizador automático de dependencias. No siempre es posible contar con recursos automáticos de esta naturaleza, ni para cualquier idioma ni para cualquier dominio. Y es importante también recordar que siempre que se trabaja con patrones estáticos, se tiene muy poca flexibilidad y cobertura en los resultados.

2.4. Oposición

Tomaremos como oposición aquellas relaciones en las que se tiene cierto grado de exclusión entre las instancias de los conjuntos dados. No es posible hablar de absolutos, por lo que es conveniente, primero, exponer las dos principales formas de encontrar oposición, siguiendo el esquema de relaciones expuesto en la figura 2-1: compatibles e incompatibles. Más adelante expondremos una segunda forma de clasificación de oposición más relevante para los propósitos de esta investigación: sistémica y no-sistémica.

2.4.1. Compatibles e incompatibles

La compatibilidad corresponde a la superposición de las relaciones que se muestran en la figura 2-1. Esta es quizá la relación menos mencionada de todas; la razón es que los conjuntos que guardan este tipo de relación aporcan muy poca o nula información sobre las entidades que participan. Es decir, la aseveración de una cualidad no implica ni restringe otra; de esta manera, casi todo el conjunto de adjetivos (con la notoria excepción de los opuestos) está relacionado de esta manera. Algo puede, por ejemplo, ser "grande" y ser "amarillo", ser "mortal" y ser "pequeño", ser "raro" y ser "redondo".

No obstante, la compatibilidad se vuelve más relevante con el caso particular en que las palabras en cuestión tengan un hiperónimo común, es decir, que sean co-hipónimos. En estos casos las palabras comparten algunos rasgos semánticos, mientras que cuentan con otros que no se contraponen del todo [Cruse *et al.*, 1986]. Un ejemplo claro es la relación que existe entre "perro" y "mascota", ambas son instancias de animales (co-hipónimos), pero no todos los perros son mascotas ni todas las mascotas son perros. Otro ejemplo se puede ver en los colores, ya que si bien como concepto los colores no son compatibles entre sí (por ejemplo, el rojo no es azul, no puede ser rojo y también azul), como propiedad sí son compatibles, ya que pueden ser aplicados simultáneamente a un mismo objeto (por ejemplo, una abeja es al mismo tiempo negra y amarilla); y precisamente es este tipo de caso es donde más se presenta una oposición compatible.

La incompatibilidad corresponde con la disyunción en las relaciones que se muestran en la figura 2-1. Es decir, no existe una instancia que mantenga cualidades de los dos conjuntos; si una aseveración para un conjunto es verdadera entonces implica que no puede ser verdadera para el otro. Por ejemplo, si X es una "lámpara", no puede ser "lluvia"; si X es un "sillón", no puede ser una "flor", etc. Del mismo modo que ocurre con los conjuntos compatibles, la incompatibilidad general no ofrece mucha información útil.

Sin embargo, al igual que con los conjuntos compatibles , la relevancia e interés en la información que aportan estas relaciones es cuando ocurren entre instancias con un hiperónimo común (co-hipónimos). En estos casos podemos observar la formación de instancias excluyentes dentro de una sola categoría, por ejemplo "animal": perro, gato, león, elefante, etc.

Como se puede apreciar, tanto en los conjuntos compatibles como en los incompatibles, el principal interés surge cuando se analizan palabras que pertenecen a un mismo campo semántico [Cruse *et al.*, 1986]; esta restricción, como se ha mencionado anteriormente, implica que las palabras sean co-hipónimos. Y esa es una implicación que seguiremos haciendo por el resto de esta investigación, no trabajaremos con toda la oposición que puede ser incluida en esta categorización de compatibilidad. Por el contrario, vamos a trabajar exclusivamente con la oposición que esta incluida dentro de aquellas palabras que estén incluidas en un mismo campo semántico. Dentro de dichas palabras, aún podemos apreciar distintos grados o categorías de oposición que, ya delimitada por lo anterior, son mucho mas relevantes para esta investigación: la oposición sistémica y no-sistémica.

Dentro de las relaciones léxicas de oposición se pueden encontrar muchas categorizaciones de acuerdo al tipo y grado que conlleve la oposición de las palabras. Se pueden distinguir, por ejemplo, los pares "opuestos" que incluyen pares como *muerto-vivo* o *recto-curvo*; este tipo de oposición es absoluta y mutuamente excluyente, no es posible mantener ambas cualidades al mismo tiempo e incluso es imposible no tener ninguna [Lyons, 1977].

Se puede hacer referencia también a los "verdaderos antónimos", como *feliz-triste*, que cuentan con dos propiedades extra [Lyons, 1977]:

- Es posible no tener ninguna de las dos cualidades.

- Son graduables, es decir, hay todo un rango en medio de los puntos extremos.

El término "verdaderos antónimos" es una herramienta para distinguirlos de los "antónimos" coloquiales, que se refieren a una combinación de los casos que se han desarrollado.

La categoría más polémica para autores como Lyons [1977] es la de "múltiples incompatibles", que incluyen principalmente categorías cerradas como las estaciones del año, lugares de origen o rangos militares. Eeste tipo de relación comparte muchas características con los "opuestos", principalmente el ser excluyentes y a veces graduables entre sus extremos (como los colores); sin embargo, la característica central que no comparten con los "opuestos" es la cualidad de ser binario.

Como se menciona en Lobanova *et al.* [2010], los estudios de relaciones léxicas convencionales solo harían referencia a este tipo de relaciones como co-hipónimas, sin tomar en cuenta su potencial de oposición.

2.4.2. Oposición sistémica y no-sistémica

Un análisis completo sobre las diferentes formas que existen de clasificar la oposición va más allá de los alcances de este trabajo. Sin embargo, para los propósitos de esta tesis se trabajará una clasificación de oposición, la cual se basará en la teoría de oposición propuesto por Mettinger *et al.* [1994]. En dicha propuesta se distinguen dos tipos fundamentales de oposición:

Sistémica: La oposición sistémica consiste en una relación estable de opuestos independiente del contexto, esto quiere decir que la oposición ha logrado anclarse en la estructura semántica del lenguaje (pares como *amor/odio*, *grande/pequeño*). Las palabras en cuestión mantienen una separación tan amplia y marcada que suelen ser utilizadas para resaltar el contraste entre otros elementos originalmente no tan diametrales. Un buen ejemplo de esto se puede encontrar en el poema 20 de Pablo Neruda[3]:

> Es tan **corto** el *amor*, y es tan **largo** el *olvido*.

[3]De Veinte poemas de amor y una canción desesperada

En dicha frase se puede ver el uso de la oposición sistémica denotada por *corto-largo* usado como herramienta para aumentar el contraste entre las ideas que se mencionan (amor/olvido).

No-sistémica: En la oposición no-sistémica, el contraste solamente se revela en determinados contextos, aquellos en los que las dos palabras aparecen juntas a modo de contraste; si bien en el significado de las palabras no existe nada que que las contraponga, no es extraño encontrar una dicotomía. Ejemplo de esto se puede ver en el par *amor/dinero*: nada evita que estas dos cualidades existan juntas, sin embargo diversos factores han llevado a la percepción de cierta limitante en su co-ocurrencia, esto se puede ver en el surgimiento de frases de origen popular como:

> Con dinero tienes a la pareja que quieras, sin dinero tienes a la pareja que te ama.

Más, aún, estas relaciones ya no son binarias, pues se pueden encontrar expresiones como la de Alberto Moravia, donde nuevamente se hace evidente una oposición no-sistémica pero ahora con un elemento distinto (amor/matrimonio):

> El amor es un juego; el matrimonio, un negocio.

Como se puede observar, son los contextos los que les dan las connotaciones de oposición, no obstante, la frecuencia de dichas connotaciones es suficiente para lograr la percepción de contraste en un gran número de pares de palabras en los que la oposición no es apreciable fácilmente, ejemplo de esto pueden ser pares como *cabeza/cola* o *teoría/práctica*.

A pesar de que el principal interés de Mettinger radicaba en la oposición sistémica, encontró que la oposición no-sistémica tiene una mayor presencia cuando se trabaja con patrones. No obstante, como se verá en la siguiente sección sobre detección de la relación, no es común utilizar estas clasificaciones de oposición en los desarrollos de extracción automática; por lo general, engloban sus estudios en términos de "antónimos" en el sentido general de la palabra.

En el desarrollo de esta investigación, se obtienen dos conjuntos de datos de salida, los cuales estaremos estudiando como candidatos a co-hipónimos y antónimos, según sea el caso, y es para dichos "antónimos" que se llevará a cabo en la sección 7.2 una evaluación sobre el grado de oposición que presentan con respecto a la clasificación que se acaba de presentar en esta sección.

2.4.3. Detección de oposición

La extracción de relaciones léxicas, no solo de oposición, es esencial para la construcción y expansión de recursos léxicos. El enfoque más común en la literatura para esta tarea es la extracción de patrones de texto no estructurado [Al-Yahya *et al.*, 2016]. Entre palabras se pueden encontrar relaciones de similitud, sinonimia, hiperonimia, contraste, antonimia, homonimia, hiponimia, etc. Pero de entre estas relaciones, las más comunes de encontrar trabajadas en la literatura son las de sinonimia, antonimia e hiperonimia, de las cuáles ya se ha hablado en sus respectivas secciones dentro de este capítulo.

La hiperonimia y la sinonimia son las relaciones más estudiadas en la literatura, pero también se pueden encontrar trabajos enfocados en la antonimia y la co-hiponimia. No obstante, también mantienen la misma tendencia de uso de patrones y conjuntos con palabras iniciales semilla, por lo regular obtenidas previamente de forma manual.

Dentro de los trabajos que buscan formas para extraer automáticamente relaciones léxicas podemos encontrar el caso de Lobanova *et al.* [2010]. En esta investigación, se usan pares antónimos semilla para detectar patrones en un corpus en idioma holandés y con ellos a su vez extraer más pares antónimos. Al llevar a cabo dicho proceso, encontraron que los mismos patrones que dan lugar a la extracción de antónimos indisputables como *pobre-rico* (antónimos que podríamos considerar con un grado de oposición sistémica) también dan lugar a la extracción de co-hipónimos que podrían ser percibidos como opuestos en el contexto correcto como *atlético-flaco* (es decir, opuestos no-sistémicos, categoría que para los fines de esta tesis también entraría en condición de antónimo).

Mettinger también describe estructuras donde es común encontrar palabras de oposición,

ventanas de contexto como *X y Z* o *ni X ni Y*, mismos patrones a los que llegan también otros autores [Justeson y Katz, 1991; Fellbaum, 1995; Jones, 2003]; este tipo de ventanas de contexto resaltan por coincidir con los patrones simétricos (ver Sección 2.4.3). Estos patrones han sido utilizados en investigaciones como Jones *et al.* [2007] para medir la estandarización de ciertos antónimos con la ayuda de corpus generales; sin embargo, no se encontraron referencias de investigaciones donde se usaran estos patrones para la extracción automática de antónimos.

Antónimos

En el caso de los antónimos, por ejemplo, Lobanova *et al.* [2010] extraen antónimos a partir de noticias en holandés con la ayuda de patrones léxico-semánticos. Los autores parten de un conjunto de antónimos bien estudiados y bien establecidos, los cuales usan para encontrar indicadores de otros antónimos, es decir, patrones. Al igual que en los otros casos, los patrones se usan a su vez para buscar nuevos pares de antónimos y el proceso se repite: pares – patrones – pares. La investigación también evalúa sus resultados con la ayuda de cinco etiquetadores humanos, gracias a lo cual encontraron que la mayoría de los pares extraídos no estaban presentes en recursos léxicos. Además, muchos de los pares que encontraron fueron descritos como co-hipónimos, y de una forma compatible a la oposición no-sistemica. De acuerdo con los autores, este tipo de pares de palabras ha sido poco estudiado en la literatura, y argumentan que esa clase de co-hipónimos deberían ser considerados como un subtipo de antónimo, que es precisamente lo que se está haciendo en este trabajo, con la consideración de la oposición no-sistémica.

Otro ejemplo de detección de antónimos lo podemos encontrar en el sistema Badea, reportado por Al-Yahya *et al.* [2016] que permite extraer de manera semi-automática relaciones léxicas de un corpus en idioma árabe. En este caso, la estrategia fue el uso de ontologías para la obtención de los antónimos semilla. A partir de allí, el procedimiento es el mismo: encontrar patrones en donde los antónimos co-ocurren, para luego usar esos patrones en la búsqueda de nuevos pares. En sus resultados, muestran la efectividad de los sistemas basados en patrones para enriquecer la información de lexicones.

En Roth y Upadhyay [2019], se detectan antónimos entre palabras con el uso de patrones especiales. En dicho trabajo, la investigación se enfoca en la distinción entre sinónimos y antónimos con la ayuda de patrones y marcadores de discurso. Los autores encuentran motivación en estudios lingüísticos que apoyan y muestran conexión entre palabras cuando hay una conexión discursiva entre ellas. Lo que encuentran en sus experimentos es que las personas por lo regular usan tipos específicos de marcadores para expresar contraste entre las palabras. Para encontrar relaciones discursivas de forma efectiva, en Roth y Im Walde [2014] los autores usaron marcadores de discurso como los que se muestran en la tabla 7-6. El método de Roth y Upadhyay [2019] se caracteriza por el uso de un módulo de traducción, con lo que pueden asociar la traducción con modelos entrenados en idiomas distintos.

En Roth y Im Walde [2014], se detectan antónimos entre palabras con el uso de patrones especiales. En dicho trabajo, la investigación se enfoca en la distinción entre sinónimos y antónimos con la ayuda de patrones y marcadores de discurso. Los autores encuentran motivación en estudios lingüísticos que apoyan y muestran conexión entre palabras cuando hay una conexión discursiva entre ellas. Lo que encuentran en sus experimentos es que las personas por lo regular usan tipos específicos de marcadores para expresar contraste entre las palabras, tales como los que se muestran en la tabla 7-6. Posteriormente, Roth y Upadhyay [2019] asocian la traducción con modelos entrenados en idiomas distintos.

Esta característica les permite aprovechar el contraste de marcadores de discurso como parte central de la traducción, ya que en el idioma objetivo de la traducción es común encontrar varias opciones para una palabra dada. Como se menciona en el artículo como ejemplo: cuando se busca en un *embedding*, las palabras “menos” y “más” comparten el mismo vecino cercano en alemán: *mehr*; y es para este tipo de casos que se buscan marcadores discursivos en el contexto como fuente de pistas que permitan una correcta identificación de las relaciones.

Este último método para la detección de relaciones logra una gran precisión y es muy confiable; sin embargo, también tiene dos desventajas importantes: la primera es la necesidad de previos estudios lingüísticos que, a menudo, tienden a encontrar patrones para corpus de propósito general, por lo que se descuidan los casos especializados; la segunda desventaja es

Tabla 2-2: Ejemplos de marcadores discursivos Roth y Im Walde [2014].

CONTRAST	but, altough, rather ...
RESTATEMENT	indeed, specifically, ...
INSTATIATION	(for) example, instance, ...

que el algoritmo es capaz de detectar la relación únicamente cuando las palabras co-ocurren exactamente en el patrón esperado.

Co-hipónimos

Los co-hipónimos han sido, en general, mucho menos estudiados que las relaciones anteriores que se han presentado, pero también podemos encontrar trabajos muy interesantes y relevantes para esta investigación.

Uno de dichos trabajos es el presentado por Jana y Goyal [2018], en donde se presenta un método supervisado para la detección de co-hipónimos. Dicha investigación se basa, a su vez, en una representación especial de grafos [Riedl y Biemann, 2013] que se construyen a partir de n-gramas sintácticos usando *Google books* como corpus. La principal hipótesis del trabajo se basa en que se asume que cuando dos palabras comparten una relación de co-hiponimia, entonces también tendrán una similitud distribucional mayor comparado a otro tipo de relaciones como la hiperonimia o la meronimia.

De forma más específica, para calcular la similitud entre dos palabras dentro de un grafo, en Jana y Goyal [2018] los autores experimentan con el uso de cinco métricas propias de los grafos: similitud estructural, camino más corto, camino más corto ponderado, densidad de conexión entre la intersección de las vecindades, y densidad de conexión entre la unión de las vecindades.

Patrones simétricos

Otro concepto importante y de gran relevancia para esta investigación es el de patrones simétricos. Los patrones simétricos no están dirigidos a encontrar antónimos o co-hipónimos en particular; de hecho, ni siquiera son presentados como un método para la búsqueda de

oposición, salvo por casos particulares, pero consideramos que son la herramienta más cercana a la metodología que se obtiene en esta investigación; se puede notar cierta oposición en los resultados que logra extraer, sobre todo en los patrones especiales que mencionaremos más adelante.

Los patrones simétricos son patrones léxicos en los que un par de palabras X y Y pueden aparecer de manera intercambiable dentro del patrón. Ejemplos de este tipo de patrones son: "X y Y" o "tanto X como Y" Este tipo de patrones se han utilizado para encontrar y representar similitud en tareas como adquisición léxica [Widdows y Dorow, 2002], agrupación de palabras [Davidov y Rappoport, 2006], clasificación de entidades nombradas [Kozareva *et al.*, 2008], identifiación de sarcasmo [Davidov *et al.*, 2010b], análisis de sentimiento [Davidov *et al.*, 2010a], atribución de autoría [Schwartz *et al.*, 2013], categorización semántica de palabras [Schwartz *et al.*, 2014] y predicción de similitud de palabras [Schwartz *et al.*, 2015].

Davidov y Rappoport [2006] desarrollaron un algoritmo no supervisado para la extracción automática de patrones simétricos. El algoritmo consiste en extraer una secuencia de 3 a 5 tokens, dentro de las cuales se puedan encontrar 2 palabras intercambiables, y de 1 a 3 palabras fijas. Se recorre todo el corpus en búsqueda de los patrones más frecuentes que coincidan con dicha plantilla. Los patrones que se encuentran se aceptan si el orden de las palabras de menor aparición es de al menos 15 % (u otro valor a asignar) con respecto al de mayor aparición.

2.5. La negación como oposición

La negación es el fenómeno que relaciona una expresión e a otra con un significado que está opuesto de alguna manera al significado de e [Horn y Wansing, 2017]. Se trata de un fenómeno que está presente en todos los idiomas y en todos ellos está marcado de forma léxica [Jiménez-Zafra *et al.*, 2020]. Se usa para revertir el significado de una parte de una oración, significado que por defecto es afirmativo; además, la negación suele tener una implicación positiva y viceversa [Blanco y Moldovan, 2011].

De los ejemplos de la sección anterior, resalta esa constante con los términos en oposición:

tienden a mantener una implicación negativa entre las palabras que están relacionando, principalmente la oposición sistémica. Es decir, si se toma como verdadero un término, por ejemplo **vivo**, esto implica la negación de su opuesto: **no muerto**.

En esencia, la implicación textual se refiere a la relación que existe entre un par de expresiones denotadas como T (el texto que va a implicar) y H (la hipótesis implicada). Se dice que T implica H si de la lectura de T típicamente se infiere que H es verdadero para un lector humano [Dagan *et al.*, 2005].

La detección de implicación en los textos es el núcleo de tareas como la inferencia léxica, la cual, a su vez, es de gran importancia para muchas aplicaciones del PLN en donde sea relevante la información implícita que conllevan los textos, como los sistemas de pregunta-respuesta, de extracción de información, de traducción automática, de paráfrasis o el resumen automático [Pekar, 2008; Bentivogli *et al.*, 2017].

Por su parte, la implicación corresponde también a un tipo de relación; sin embargo, a diferencia de las relaciones léxicas (ver sección 2.1), la implicación se da sobre todo a nivel enunciado, más que a nivel palabras. No obstante, se puede apreciar una gran coincidencia entre la implicación léxica y las relaciones léxicas, ya que la implicación por lo regular se concentra en un conjunto reducido de palabras dentro de los enunciados. En este sentido, se puede observar, por ejemplo, que muchos enunciados pueden quedar implicados debido a la relación hipónimo-hiperónimo:

- Todo **individuo** tiene derecho a la vida, a la libertad y a su seguridad personal. → Toda **mujer** tiene derecho a la vida, a la libertad y a su seguridad personal.

- Isabel II es la actual **reina** británica. → Isabel II es una **mujer**.

En el primer ejemplo del enunciado anterior, tenemos una implicación dada por un hipónimo, mientras que en la segunda tenemos una implicación dada por un hiperónimo. Este par de enunciados también reflejan cómo, aunque están vinculados, las relaciones léxicas y la implicación no mantienen una relación unívoca y depende del contexto.

Como ya se mencionó anteriormente, a diferencia de los hipónimos o hiperónimos, la oposición está mucho más relacionada con una implicación negativa y, por tanto, puede ser aprovechada de forma particular cuando existe negación en los textos. Este tipo de implicación ha sido utilizado para buscar formas de representar la negación a través de las implicaciones positivas que conlleva [Blanco y Moldovan, 2011].

En particular, se puede observar un gran interés en la detección de la negación para la tarea de minería de opinión. Shah y Rekh [2014] analizan el aumento de popularidad de las redes sociales, medios que se presentan como una forma versátil e instantánea de comentar sobre cualquier tema, producto o servicio; esto le da a las redes sociales un carácter muy ideal para la minería de opiniones, y es la razón por la que se enfoquen tantos esfuerzos en esta tarea.

El objetivo principal de la minería de opinión es la extracción automática de opiniones expresadas en un texto dado; la opinión es un concepto amplio en el que se involucra juicio humano, valoración de normas y conductas sociales, el entendimiento de desempeño esperado de productos, servicios, actitudes o imágenes estéticas, etc.

Se trata de una tarea de gran dificultad, principalmente por la ambigüedad que se tiene en la granularidad [Shah y Rekh, 2014]: La tarea de análisis de sentimiento se puede hacer para encontrar documentos relevantes, para encontrar secciones relevantes, para encontrar el sentimiento de forma general de un texto o para cuantificar el sentimiento. Para algunos de estos cometidos se tienen que tomar en cuenta también el alcance que tiene una declaración, ya sea positiva o negativa; es decir, sobre qué objetos se están expresando las opiniones. Por ejemplo, si tenemos una oración como la siguiente:

> El celular tiene una buena calidad de llamada, pero su batería dura poco.

Nos encontramos con dos opiniones encontradas, la primera, positiva, hacia la calidad de la llamada; mientras que la segunda, negativa, es hacia la batería. En estos casos puede ser importante distinguir no solo la polaridad de las opiniones, sino los elementos a los cuáles están dirigidas. Es por ello que en minería de opinión es común encontrar sistemas de etiquetado en forma de tuplas, donde uno de los elementos es sobre quién se está opinando [Liu, 2012].

La negación es una construcción lingüística muy común que se comporta de forma parecida, es decir, también tiene un área de efecto específica (*scope*). Además, afecta la polaridad, por eso es un elemento de mucho interés para la tarea de minería de sentimientos [Shah y Rekh, 2014]. La negación se puede encontrar en un texto principalmente por indicadores léxicos específicos como: no, ni, ningún, etc. [Bel-Enguix *et al.*, 2021]. Por ello, existen algunas estrategias simplistas para el manejo de la negación, por ejemplo tomar los sustantivos o verbos mas cercanos a la negación para ser tomados como *scope*. Sin embargo, se ha demostrado que las expresiones negativas no implican negación, y la tarea de la detección y delimitación del *scope* es muy compleja; para poder aprovechar la información de la negación es necesario desarrollar sistemas especializados en dicha tarea [Na *et al.*, 2005; Fancellu *et al.*, 2016]. La detección de negación, en particular en español, es complicada debido a que los marcadores de negación puede ser discontinuos y ambiguos; además, los *scopes* pueden aparecer ya sea antes o después de los marcadores [Jiménez-Zafra *et al.*, 2020]. La tarea se vuelve aún mas compleja cuando se analizan textos de redes sociales; sin embargo, es importante poder analizar estos últimos para obtener opiniones heterogéneas de diversas áreas de interés [Bel-Enguix *et al.*, 2021].

No obstante, detectar la negación y su área de efecto no es todo lo que se necesita para poder aplicar información de negación en tareas como la minería de opinión. También es necesario encontrar una representación de esa negación sobre su *scope*. Y es en esa etapa en la que se pueden utilizar los términos de oposición. Ya hemos desarrollado la relación que existe entre la oposición y la negación. Si se cuenta entonces con pares de oposición, se puede encontrar un proceso de transformación entre los términos opuestos que, mas adelante, sea aplicable sobre elementos negados (aquellos que formen parte del *scope* de una negación).

En las sección 7.4 ejecutamos un experimento sobre la aplicación de minería de opinión para evaluar los efectos que puede lograr una transformación de oposición a partir de los resultados de esta investigación. Como parte de la anotación de negación, nos basamos en el trabajo de Bel-Enguix *et al.* [2021].

Capítulo 3

Un punto ciego en las semántica distribucional

Dentro del área de procesamiento de lenguaje natural, el surgimiento de técnicas capaces de aprovechar la semántica distribucional ha marcado una explosión en el avance de todas las tareas del PLN, son técnicas y modelos tan importantes que es necesario asignarles un capítulo propio para su análisis. Sin embargo, existe una vínculo particular entre la semántica distribucional y las relaciones léxicas que hemos visto anteriormente. Este capítulo proporcionará no solo una introducción a este tipo de sistemas, sino también la forma en que pueden y han sido aprovechadas las relaciones léxicas para enriquecer modelos ya de por si tan poderosos.

En la sección 2.2 se mencionaban ya dos grandes categorías en los métodos de extracción de sinónimos. Por un lado, se pueden usar patrones para la extracción y, por el otro, se pueden usar técnicas para obtener un espacio vectorial en donde cada palabra se puede representar como un punto o vector; palabras parecidas estarán cerca entre sí y eso funciona para la obtención de sinónimos, o por lo menos para palabras muy parecidas semánticamente.

Recordamos esto, ya que es interesante notar que de las relaciones léxicas que se estudiaron, los sinónimos son los únicos en los que se usa este tipo de técnica para su extracción directa. Esto responde a una propiedad intrínseca de la semántica distribucional y a la forma en la que calcula la asociación entre las palabras. En contraste con la forma en que los sinónimos son la

relación con mayor compatibilidad para la asociación de la semántica distribucional, la oposición es la relación que a estos sistemas más les cuesta distinguir, y esta dicotomía es la razón por la que las representaciones distribucionales juegan un papel tan importante en esta investigación, o puesto de otro modo, es la razón por la que esta investigación tiene tanto potencial para su aplicación en el enriquecimiento de representaciones distribucionales.

3.1. Hipótesis distribucional

"A word is characterized by the company it keeps" [Firth, 1957]; una palabra está caracterizada por la compañía que mantiene, es la esencia de la hipótesis distribucional y una de las principales fuentes de información que se usan para la obtención de representaciones vectoriales de las palabras.

El principio detrás de la hipótesis distribucional se puede visualizar fácilmente utilizando una analogía de la visión: Cuando se tiene una imagen, y se divide en pequeñas secciones, una pequeña sección tendrá información como color, luminosidad, etc; pero es muy difícil entender la función del fragmento dentro de la imagen completa sin contar con la información que lo rodea. Esto se puede apreciar fácilmente, ya que incluso ha dado lugar al desarrollo de juegos que consisten en intentar adivinar el objeto de una imagen luego de que se le ha hecho un acercamiento a una zona específica [1]. En lenguaje natural ocurre algo similar, también se tienen significados que se van formando en donde cada fragmento juega un papel (podemos enfocarnos en las palabras para este fin) que da información por separado, pero su interpretación completa depende de su contexto.

De manera más formal, esto ha sido ampliamente estudiado, siendo uno de los trabajos mas reconocidos la hipótesis distribucional de [Harris, 1954, 146], quien afirma que un elemento puede ser caracterizado al tomar en cuenta todos los contextos en que puede aparecer, para lo que da la siguiente definición de contexto: "An environment of an element A is an existing array of its co-ocurrences, i.e. the other elements, each in particular position, with which A occurs to yield an utterance". Dicho de otra forma, dos instancias x y y se pueden incluir en un elemento

[1] `https://www.youtube.com/channel/UCDFHVL6VQu57DpElOaMYWOQ`

A si la distribución de x relativo a otros elemenos (B, C, etc.) es similar a la distribución de y.

Si bien los procedimientos distribucionales de Harris [1954] fueron originalmente introducidos para análisis fonológico, se han convertido en una metodología general aplicada aplicada a muchos niveles de la lingüística [Goldsmith, 2005]. Podemos encontrar trabajos que buscan aprovechar el núcleo de estos mismos planteamientos, desde análisis a nivel morfológico como a nivel documentos dentro de temáticas similares [Mijangos de la Cruz, 2015].

Para sistemas automáticos, utilizar esta información representa una ventaja importante, debido a que supone la posibilidad de entrenar modelos para codificar significados de palabras en vectores de manera no supervisada, es decir, sin la necesidad de corpus anotados.

3.2. *Word embeddings*

Comprender el significado de una palabra en su contexto es el corazón del procesamiento del lenguaje natural, y en la búsqueda de este fin, uno de los desarrollos mas importantes que han tenido lugar es la representación de las palabras en espacios vectoriales densos; dentro de esta estrategia, muchos métodos han tenido éxito en el modelado de ciertos aspectos de similitud entre palabras, principalmente los que parten del uso de la semántica distribucional desarrollada en la sección anterior; estos sistemas normalmente obtienen representaciones vectoriales para las palabras a partir de sus contextos y las distribuciones de frecuencias que rodean a dichas palabras.

Dentro de las representaciones distribucionales, las más representativas son los llamados *word embeddings*, modelos de espacios vectoriales (VSM por sus siglas en inglés) que representan palabras como vectores de valores reales en un espacio semántico de baja dimensión (relativo al tamaño del vocabulario) [Iacobacci *et al.*, 2015].

En la práctica, para poder aprovechar los contextos de las palabras, la mayoría de los sistemas toman modelos de lenguaje como punto de partida para llevar a cabo el entrenamiento del modelo. Una de las principales tareas de un modelo de lenguaje es ser capaz de predecir la siguiente palabra dentro de una secuencia, para lo cual se podrían usar las cuentas de las

secuencias de palabras con respecto a secuencias incompletas [Ruder, 2016a]:

$$p(w_t|w_{t-1},\ldots,w_{t-n+1}) = \frac{count(w_{t-n+1},\ldots,w_{t-1},w_t)}{count(w_{t-n+1},\ldots,w_{t-1})} \tag{3-1}$$

Donde w_t representa la palabra objetivo que se busca aprender a predecir y n la longitud de la secuencia. Obtener cuentas para cadenas muy largas se vuelve impráctico, por lo que se recurre a la suposición de Markov y tomar en cuenta solo las palabras más cercanas (valores pequeños de n).

Las representaciones distribucionales parten de este tipo de modelo, pero tienen una diferencia clave: no se ven limitadas únicamente al conocimiento de las palabras previas de una palabra objetivo; es decir, si por un lado los modelos de lenguaje buscan predecir la última palabra a partir de las anteriores, los modelos distribucionales buscan, por su parte, predecir una palabra intermedia a partir de las que la rodean; esto permite poder tomar en cuenta palabras posteriores y no solo previas. De esta manera, los modelos distribucionales se pueden entrenar, teóricamente, buscando maximizar una función objetivo (J), compuesta por la probabilidad conjunta de cada una de las palabras objetivo w_t dadas las palabras de sus contextos, según un corpus de entrenamiento:

$$J = \prod_{\substack{t\in V \\ c\in\text{contexto}}} P(w_t|w_c) \tag{3-2}$$

En donde V es el vocabulario del corpus. En la práctica, es preferible utilizar los logaritmos de las probabilidades de la palabra objetivo w_t para una ventana de contexto acotada a una longitud l, ya que los productos se convierten en sumas y es más factible el cómputo. Así, la función objetivo (J) de la ecuación 3-2 puede ser transformada a su versión logarítmica ($\hat{J}$) de la siguiente manera:.

$$\hat{J} = \sum_{t\in V} \log P(w_t|w_{t-l},\cdots,w_{t+l}) \tag{3-3}$$

$$\hat{J} = \sum_{t \in V} \sum_{-l \leq c \leq l, c \neq 0} \log P(w_t | w_{t+c}) \tag{3-4}$$

Para la tarea de predicción, las palabras son representadas como vectores y por lo regular se hace uso de un clasificador multiclase como *softmax*; con lo que tenemos [Mikolov *et al.*, 2013a]:

$$P(w_i | w_j) = \frac{e^{\bar{w}_i^\top \bar{w}_j}}{\sum_{k \in V} e^{\bar{w}_i^\top \bar{w}_k}} \tag{3-5}$$

$$\hat{J} = \sum_{t \in V} \sum_{-k \leq c \leq k, c \neq 0} \log \frac{e^{\bar{w}_t^\top \bar{w}_{t+c}}}{\sum_{k \in V} e^{\bar{w}_t^\top \bar{w}_k}} \tag{3-6}$$

Debido a la complejidad y al costo computacional necesario para calcular los vectores que satisfacen las funciones, se han llevado a cabo investigaciones para optimizar los algoritmos de entrenamiento, mismas que han dado lugar a la creación de sistemas sumamente eficientes como *word2vec* y *GloVe* [Mnih y Hinton, 2009; Morin y Bengio, 2005; Pennington *et al.*, 2014; Mikolov *et al.*, 2013a].

Este tipo de modelos predictivos han tenido un gran éxito y se han popularizado dentro de la comunidad de PLN. Baroni *et al.* [2014] se dan a la tarea de realizar una serie de evaluaciones para comparar los nuevos modelos contra los enfoques clásicos, cuyos vectores se basaban en conteos y llegan a la conclusión de que los *word embeddings* se han ganado de manera justificada la popularidad que ahora disfrutan.

El enfoque de *word embeddings* sostiene que el significado de las palabras debe estar representado por la distribución que tengan las palabras dentro de un espacio vectorial. Es decir, que una palabra sea identificada con una posición en un espacio vectorial y que las representaciones vectoriales de palabras similares se encuentren cercanas entre sí. Esto es, por ejemplo, que la distancia entre la representación vectorial de "árbol" y "arbusto" sea mucho menor que la distancia entre "árbol" y "computadora". Estas representaciones vectoriales del significado de palabras aprovechan la información de co-ocurrencia de pequeñas ventanas, junto con medidas como coseno para medir la distancia entre ellas, y esto ha resultado ser la mejor representa-

ción semántica en diversas tareas como búsqueda de sinónimos y categorización sintáctica y semántica [Bullinaria y Levy, 2007].

Los *word embeddings* se han utilizado en una gran variedad de tareas, incluyendo la inducción de léxico bilingüe [Mikolov *et al.*, 2013b], el análisis de sentimientos [Melamud *et al.*, 2016; Socher *et al.*, 2013; De Boom *et al.*, 2016], el reconocimiento de entidades nombradas [Melamud *et al.*, 2016; Turian *et al.*, 2010; Guo *et al.*, 2014], clasificación de documentos [Kiela *et al.*, 2015], rastreo de estado en diálogos [Mrkšić *et al.*, 2016], análisis de dependencias [Melamud *et al.*, 2016] resolución de correferencia [Melamud *et al.*, 2016].

El éxito que han tenido los *word embeddings* se debe en gran medida a que son modelos muy eficientes al ser entrenados en corpus de datos sumamente grandes y a que se les ha considerado representaciones vectoriales generales del significado de las palabras.

Word2Vec

En 2013 se presenta el que es probablemente el modelo mas popular de *word embeddings: word2vec* [Mikolov *et al.*, 2013b,a]. Quizá la principal aportación que se logra en el desarrollo de dichos *word embeddings* radica en el entrenamiento eficiente que se logra para poder aprovechar grandes cantidades de datos textuales. Mikolov *et al.* [2013a] proponen el método *skip-gram* para el entrenamiento de vectores de palabras a partir de un corpus. El objetivo estándar del método está dado por la ecuación:

$$\hat{J} = \sum_{\substack{i \in corpus \\ j \in contexto(i)}} \log \left(\frac{\exp\left(w_i^\top \tilde{w}_j\right)}{\sum_{k=1}^{V} \exp\left(w_i^\top \tilde{w}_j\right)} \right) \tag{3-7}$$

donde las w_i son las representaciones vectoriales de la palabra i y las $\tilde{w}_j$ son las representaciones vectoriales de las palabras de contexto j

La ecuación anterior es equivalente a la ecuación 3-6. Como se puede observar en ambos casos, cada uno de los elementos para el *softmax* requiere un producto entre las representaciones de la palabra objetivo y todas las palabras del vocabulario para obtener la suma del denominador y obtener la probabilididad normalizada de la palabra objetivo en su contexto.

Para lidiar con la complejidad computacional de la suma del denominador antes mencionada, Mikolov *et al.* [2013a] usan una estrategia conocida como *Negative Sampling*, en la que se usa una función de pérdida auxiliar que también busca optimizar el objetivo de maximizar la probabilidad de palabras correctas. En esencia, se entrena un sistema cuyo objetivo es clasificar conjuntos de datos correctos en contraste con versiones "corruptas" de ventanas de contexto en las cuales fueron insertadas palabras aleatorias [Ruder, 2016b].

GloVe

Otro sistema de *word embeddings*, casi igual de popular que *word2vec*, se presenta en *GloVe*, el sistema que se va a estar utilizando en las evaluaciones de enriquecimiento de esta investigación. Pennington *et al.* [2014] presentan un sistema de *embeddings* que se obtienen a partir de un objetivo por mínimos cuadrados sobre una matriz de co-ocurrencia global X, en la que X_{ij} es el número de veces que aparece la palabra j en el contexto de la palabra i:

$$\hat{J} = \sum_{\substack{i \in corpus \\ j \in contexto(i)}} f(X_{ij})(w_i^\top \tilde{w}_j - \log X_{ij})^2 \tag{3-8}$$

donde:

$$f(x) = \begin{cases} (x/x_{max})^a & \text{si } x < x_{max} \\ 1 & \text{de lo contrario} \end{cases}$$

Donde a y x_{max} son hiperparámetros que, siguiendo a los autores, se establecen en $a = 3/4$ y $x_{max} = 100$.

Con esta nueva función objetivo, *GloVe* busca lograr de forma explícita una cualidad que *word2vec* obtiene parcialmente de forma fortuita [Ruder, 2016c]; específicamente, Pennington *et al.* [2014] sostienen que es el radio de las probabilidades de la co-ocurrencia de palabras, y no la propabilidad de las co-ocurrencias, lo que contiene la información de interés para codificar en un espacio vectorial.

Tanto *Word2Vec* como *Glove* se encuentran dentro de los sistemas base que inspiran y provo-

can el desarrollo acelerado de otras técnicas de *embeddings*. Por ejemplo, *fastText* [Bojanowski *et al.*, 2017] usa fragmentos menores a las palabras para encontrar codificaciones de información extra (como los afijos) que se pueda usar para extrapolar mejores representaciones de palabras no vistas anteriormente. También se pueden mencionar ejemplos de *embeddings* más actuales, como *BERT* [Devlin *et al.*, 2018], el cual lleva las representaciones vectoriales más allá y genera un modelo de lenguaje completo que permite obtener representaciones de contextos completos con representaciones de palabras dependientes del contexto en el que se están usando. Sin embargo, una revisión exhaustiva de los sistemas de *embeddings* escapa de los alcances de esta tesis; no obstante, es importante comentar que todos los sistemas antes mencionados siguen compartiendo, como principal fuente de información, los contextos de los textos que reciben como entrada.

3.3. Limitaciones

No obstante todas sus ventajas, los *word embeddings* aún tienen un largo camino por delante, se han encontrado varias limitaciones, lo que complica su uso pleno en distintas aplicaciones. Un ejemplo de las limitaciones más comunes es que la representación individual de las palabras no implica un método natural para la representación de fragmentos o textos completos [Mijangos *et al.*, 2017]; otro ejemplo que se ha abordado también es el requerimiento de grandes cantidades de texto para lograr buenos modelos [Bel-Enguix *et al.*, 2019]. Por esta razón, muchos estudios recientes se han enfocado en desarrollar métodos para optimizar y/o especializar los vectores de palabras según su aplicación.

De forma particular la limitante que se estará abordando en esta investigación es el la tendencia que tienen los *embeddings* a agrupar todo tipo de asociación como cercanía vectorial. La generalidad de significado capturada en los *embeddings* puede llegar a ser un arma de doble filo. Dicha generalidad se da debido a que las representaciones distribucionales de palabras, en su mayoría, dependen de la correlación de aparición entre palabras. Es decir, se basan en que palabras similares aparecen en contextos similares. Y si bien esto es cierto, no se puede negar

que no solo palabras similares aparecen en contextos similares.

Se puede argumentar que esta concepción de representación general puede desembocar en comportamientos no deseados en ciertas tareas del PLN. Esto debido al concepto de "similitud" entre palabras, ya que los *embeddings* no hacen ninguna distinción del tipo de similitud que guardan las palabras. No es lo mismo palabras con significado similar (carro, automóvil), que palabras asociadas (carro, gasolina); además, dentro de las palabras asociadas que se pueden encontrar en contextos similares, es igualmente factible encontrar oposición. En este sentido, los sistemas convencionales de *word embeddings* no hacen ninguna distinción entre este tipo de relaciones. Kiela *et al.* [2015] ponen como ejemplo la tarea de clasificación de documentos según su tema y afirman que es más útil tener información de palabras asociadas como "perro" y "gato" (palabras que dentro según nuestro marco de referencia lingüísticio podrían corresponder a opuestos no-sistémicos) para encontrar un tema, que tener información de sinónimos, como "perro" y "canino".

En el área de ciencias cognitivas, la diferencia entre similitud y asociación es muy clara. La similitud se da entre palabras con significados semejantes, como "gato" y "felino", mientras que la asociación se da entre palabras con significados diferentes, pero que se encuentran relacionadas entre sí, como "coche" y "gasolina". Los *word embeddings* capturan a la par tanto similitud como asociación razonablemente bien, pero en ninguno de los dos casos el resultado puede ser excelente, ya que, según Hill *et al.* [2014], los objetivos son mutuamente incompatibles; por esta razón, proponen un sistema de *word embeddings* especializados capaces de capturar, en representaciones vectoriales independientes, similitudes y asociaciones entre palabras.

Los antónimos también son un buen ejemplo de las limitaciones que presentan los sistemas distribucionales, ya que son palabras que tienen significados opuestos, que en su mayoría van a aparecer en contextos muy parecidos. Se puede reconocer la similitud que de hecho guardan los antónimos entre sí, pero existen muchas aplicaciones en el que distinguir este tipo de palabras es sumamente importante, y es algo que no se puede lograr utilizando coocurrencias de contextos como única fuente de información.

Este tipo de observaciones ha dado lugar a investigaciones [Melamud *et al.*, 2016; Mrkšić

et al., 2016; Schwartz *et al.*, 2016] que buscan la manera de enriquecer a los *word embeddings* mediante la consideración de información extra que aportan relaciones estructuradas entre palabras, como ontologías o normas de asociación de palabras[2] (más detalles de éstas investigaciones se discuten en la sección 3.4.2).

Por su naturaleza, los *word embeddings* cuentan con la cualidad de no requerir más allá del texto sobre el que se buscan los patrones y modelos del lenguaje, esto quiere decir que no se necesitan etiquetas ni estudios o herramientas previas. Esta independencia de recursos externos es una de las características que le da más fuerza y flexibilidad a los *word embeddings*. Algo que se debe tener muy en cuenta es que el proceso de enriquecimiento y especialización de los sistemas son, en su mayoría, casos que implican dejar de lado la versatilidad e independencia propias de los *embeddings*. Esto es, muchas de las herramientas modernas de PLN dependen de recursos lingüísticos para algunas de las etapas, un ejemplo de lo anterior es el uso de ontologías o tesauros [Kiela *et al.*, 2015; Abdalgader, 2016; Chen *et al.*, 2014; Chen y Manning, 2014; Ferrero *et al.*, 2017], estructuras de oraciones [Erk y Padó, 2008; Finkel *et al.*, 2005; Angeli *et al.*, 2015] o etiquetadores automáticos [Saha *et al.*, 2016; Johansson y Pina, 2015; Iacobacci *et al.*, 2015] (una descripción más detallada se puede encontrar en la sección 3.4.2). Estos enfoques desembocan en el desarrollo de sistemas altamente dependientes del idioma y de las herramientas presentes para el mismo, un resultado que en este trabajo se busca evitar en la medida de lo posible.

Considerando el planteamiento antes expuesto, se reitera la importancia de diseñar métodos generales que no requieren de recursos costosos para su funcionamiento.

3.4. Integración de relaciones léxicas

Cuando se trata de relaciones léxicas, no se puede dejar fuera el concepto de la similitud entre palabras. Sin lugar a dudas, una de las formas más populares hoy en día para buscar

[2]Las normas de asociación de palabras reflejan la relación existente entre dos palabras por parte de hablantes de una misma lengua. La técnica comúnmente empleada es la de asociación libre, la cual consiste en pedirle a los sujetos que generen espontáneamente, ya sea de manera verbal o escrita, la primera respuesta que venga a su mente al estar expuestos a una palabra estímulo [Gómez-Adorno *et al.*, 2019]

similitud entre palabras es mediante el uso de *word embeddings* (ver sección 3.3 para mayor detalles).

Levy *et al.* [2015] ejecutan una serie de experimentos para explorar las relaciones léxicas que son capaces de capturar las representaciones distribucionales de las palabras. Se centran en relaciones de hiperonimia y reportan que, según sus experimentos, a pesar de que los sistemas distribucionales que desarrollan logran buenos desempeños en la clasificación de hiperónimos, en realidad no es porque sean capaces de capturar dichas relaciones entre un par de palabras. En su reporte explican que para la búsqueda de hiperónimos, lo que captura la representación es, en realidad, la cualidad de cada palabra de qué tanto puede ser considerada, o no, como un hiperónimo en general, más que encontrar la relación de hiperonimia con sus hipónimos en particular. Este fenómeno desemboca en que el sistema genere un modelo de clasificación que busque aquellas palabras que sean más probables de tener hipónimos; estas son identificadas como "hiperónimo prototípico".

Algo que nos parece bastante interesante de los resultados que se obtuvieron son las características principales de las que se sirven sus sistemas para categorizar a las palabras como hiperónimos prototípicos. Por ejemplo, los autores hacen notar cómo muchas de las palabras dentro su corpus médico tienen como hiperónimo "*symptom*". Entonces, la principal característica discriminadora es $psychosomatic_{-1}$ (psicosomático). Esta característica, que se indica con el subíndice -1, denota que esa palabra está presente en un lugar a la izquierda de la palabra, es decir, que en el corpus se va a presentar: *psychosomatic symptom*. Para este caso, la palabra *psychosomatic* es el mayor indicador para clasificar a *symptom* como un hiperónimo prototípico. Y lo mismo ocurre para otras palabras que quedan categorizadas independientemente como hiperónimos prototípicos mediante características como: $any_{-1,-2}$, $every_{-1}$, $kinds_{-2}$, $other_{-1}$, $such_{-2,+1}$ o $including_{+1}$. Además de lo anterior, se percataron que cuando se trata de otro tipo de relaciones que no son hiperónimos, (por ejemplo relaciones cualitativas propias de una palabra [$coat \rightarrow warmth$] o componentes [$chair \rightarrow legs$]) existen otras palabras características comunes como: of_{+1} o $their_{-1}$.

En sus observaciones destacan que estos fenómenos pueden ser una limitante de las caracte-

rísticas de los sistemas distribucionales, ocasionados debido a que características como: $such_{+1}$ y $such_{-2}$, típicas para un hiperónimo prototípico y, y un hipónimos prototípico x, respectivamente (y *such as* x). En estos casos, afirman, las características contextuales de las palabras x y y no son capaces de capturar la ocurrencia conjunta de x y y en ese patrón específico, sino que se toman como características independientes para cada uno de los vectores de las palabras.

Más adelante, Roller y Erk [2016] encuentran que, si bien las aseveraciones anteriores son válidas, los sistemas distribucionales de todas maneras son una base muy fuerte para los sistemas de hiperónimos. Atribuyen esta propiedad a los vectores de contextos, más que a los vectores de palabras. Los vectores distribucionales se obtienen a partir de una matriz M en la que se concentran las coocurrencias de los diferentes contextos sintácticos:

$$M \approx WC^T$$

Al descomponer la matriz, usando *Singular Value Decomposition*, se producen las matrices W y C, de los *embeddings* y los contextos, respectivamente [Levy y Goldberg, 2014b]. Los autores defienden que es gracias a esa matriz C que los sistemas aprenden no cuáles palabras son hiperónimas, sino cuáles contextos son indicadores de hiperonimia. Contextos que, de hecho, notan que coinciden en gran medida con los patrones de Hearst [1992].

Roller y Erk [2016] además plantean una nueva representación concatenada para que se tome en cuenta también la relación de la combinación de palabras. Sin embargo, eso se logra mediante mecanismos supervisados para el aprendizaje de relaciones de hiperonimia. De una manera general, los sistemas distribucionales no son capaces de hacer aportaciones especiales de inferencia léxica.

3.4.1. Inferencia léxica

La inferencia léxica se refiere al reconocimiento de relaciones entre palabras, tales como la causalidad (gripa $->$ fiebre) y otras nociones que permitan identificar implicatura. La inferencia léxica es utilizada principalmente en tareas de similitud de palabras y para detectar

la implicación de oraciones. Dentro de las relaciones más importantes para esta tarea está la hiperonimia, que es un caso muy claro de implicación (todos los perros son animales), pero no son las únicas relaciones léxicas relevantes para la tarea.

Lenci [2008] ofrece un ejemplo interesante para los fines de esta investigación, ya que muestra que en las oraciones:

1. Google compró una nueva compañía → Google adquirió una nueva compañía.

2. John conduce un carro → John conduce un vehículo.

En ambos casos, la oración de la izquierda implica a la oración de la derecha. En la oración 1, la inferencia se debe a que las palabras "compró" y "adquirió" pueden ser considerados como sinónimos. Mientras que en la oración 2, la inferencia se debe a que la palabra "vehículo" es un hiperónimo de "carro". Y esta diferencia se vuelve más importante con los siguientes ejemplos que ofrece:

1. Google adquirió una nueva compañía → Google compró una nueva compañía.

2. John conduce un vehículo $*$ → John conduce un carro.

3. John conduce un carro $*$ → John conduce una camioneta.

En estos casos, la oración 1 sigue siendo verdadera, ya que la sinonimia es simétrica. Sin embargo, la oración 2 ahora es falsa, debido a que la relación de hiperonimia no es simétrica. Y más aún, la oración 3, en la que se presenta una relación de cohiponimia, también es falsa. "Carro" y "camioneta" son palabras semánticamente muy relacionadas y se espera que tengan contextos y por ende vectores muy similares, es por esto que los sistemas distribucionales no son capaces de capturar relaciones de inferencia léxica. Los autores están convencidos de que, en este sentido, los sistemas distribucionales están en gran desventaja contra redes semánticas como *WordNet*.

Del ejemplo anterior, se puede encontrar cierto parecido en los datos que los sistemas distribucionales no alcanzan a capturar (discutidos en la sección 3.3); es decir, relaciones de palabras que se pueden catalogar como opuestos no-sistémicos (ver sección 2.4.2 para más detalles), en el

sentido que se pueden ver elementos diferentes que forman parte de un mismo grupo semántico, pero excluyente (palabras como “carro”, “camioneta”, “motocicleta”, entre otras).

3.4.2. Similitud y asociación

Como se ha referenciado en la sección 3.3, en el área de ciencias cognitivas se hace una diferenciación clara entre similitud y asociación. La similitud se da entre palabras con significados semejantes, como “gato” y “felino”, mientras que la asociación se da entre palabras con significados diferentes, pero que se encuentran relacionadas entre sí, como “coche” y “gasolina”. Kiela *et al.* [2015] demuestran las ventajas que se pueden obtener si se hacen distinciones de estas relaciones y se especializan sistemas para uno u otro fin, ya sea para encontrar palabras relacionadas o para encontrar, la que llaman, “auténtica similitud”. En este contexto, la detección de oposición puede ser utilizada también como fuente de información que distinga dicha “auténtica similitud” de palabras opuestas que ocurran en contextos similares.

Para obtener la distinción antes mencionada, Kiela *et al.* [2015] desarrollan espacios semánticos especializados, tanto para similitud como para asociación. Para lograr la especialización de similitud utilizan el tesauro *MyThes* del proyecto *OpenOffice.org*[3], mientras que para lograr la especialización en asociación utilizaron las normas de asociación libre de la *University of South Florida (USF)*. Los *embeddings* que obtienen logran mejoras en tareas de similitud de palabras, detección de sinónimos y clasificación de documentos.

Uno de los trabajos más cercanos a la presente investigación es el de Schwartz *et al.* [2015], donde presentan un sistema de *embeddings* basado en patrones simétricos (patrones léxicos que relacionan palabras como: “X y Y” o “tanto X como Y”. Se pueden ver mas detalles en la sección 2.4.3). Con esta adición, logran mejorar los resultados de *word2vec* en la tarea de similitud de *SimLex999*, principalmente para la similitud entre verbos. Mencionan que una de las principales ventajas que presentan este tipo de enfoques es la posibilidad de controlar la manera en la que el sistema maneja los diferentes patrones, y así controlar cómo se toman en cuenta ciertas relaciones entre palabras. Por ejemplo, muestran que con ciertos patrones es

[3] `https://www.openoffice.org/lingucomponent/thesaurus.html`

posible encontrar pares de palabras antónimas que el sistema puede detectar y asignarles una métrica especial. Afirman que este tipo de modificaciones pueden ser muy importantes para tareas como clasificación de palabras y análisis de sentimientos.

En ese sentido, Schwartz *et al.* [2016] utilizan como evaluación de *embeddings* la pruebas intrínsecas de predicción humana de juicios de relación léxica para relaciones semánticas y de asociación. En particular, demuestran que los patrones simétricos logran beneficios importantes al detectar similitud de verbos y adjetivos, en contraste con sistemas de *embeddings* convencionales que obtienen muy buenos resultados principalmente en relación de sustantivos. En sus experimentos, además, comparan el desempeño de los patrones simétricos con el uso de dependencias y con el uso tradicional de ventanas de contexto. Lo que concluyen de esto es que los patrones simétricos logran obtener mejores resultados, incluso cuando las relaciones de dependencias a comparar son las de coordinación, que están muy relacionadas con los patrones sintácticos de por sí. De manera interesante, observan que tomar en cuenta los contextos de dependencias exclusivamente de coordinación supera al sistema que toma en cuenta todas las dependencias.

Mrkšić *et al.* [2016] muestran una diferencia significativa entre vectores pre-entrenados mediante *GloVe* (ver sección 3.2). Utilizan listas de sinónimos y antónimos para alejar entre sí palabras con significados opuestos, ya que, por naturaleza de los sistemas de vectorización, este tipo de palabras también se agrupa como palabras relacionadas. Con esta estrategia logran mejorar resultados en las tareas de *SimLex-999*. En su trabajo mencionan lo importantes que son para la tarea de rastreo de estado en diálogo los diccionarios semánticos como una herramienta para distinguir claramente entre palabras que de otra manera podrían considerarse muy cercanas (como el caso de "costoso" *vs* "barato" o "China" *vs* "India").

Por su parte, Melamud *et al.* [2016] proponen una variante del modelo *skip-gram* a partir de palabras candidatas a sustitución, en lugar de contextos de las palabras. Esto quiere decir que, en vez de utilizar los vectores de las palabras de un vecindario alrededor de una palabra objetivo, se emplean palabras que tengan el potencial de usar el mismo lugar que está utilizando la palabra objetivo al mantener el contexto fijo. La obtención de *word embeddings*, usando contextos por

sustitución, ha demostrado que provoca un espacio vectorial en el que se da preferencia a la similitud funcional de las palabras sobre una similitud tópica. Los contextos por sustitución se obtienen a partir de modelos de lenguaje y se han usado para obtener modelos de inducción de POS [Yatbaz *et al.*, 2012], inducción de sentido de las palabras [Başkaya *et al.*, 2013], similitud de funcionalidad semántica [Melamud *et al.*, 2014] y tareas de sustitución léxica [Melamud *et al.*, 2015]. En sus experimentos, Melamud *et al.* [2016] encontraron que los *embeddings* por sustitución, a pesar de no tener muy buenos resultados de evaluación intrínseco, ofrecen mejoras en las tareas de análisis de sentimientos y parseo de dependencias cuando se usan en combinación con otros vectores.

3.4.3. Embeddings simétricos

Por su parte, los patrones simétricos (ver sección 2.4.3) también se usan para aportar información útil para los sistemas de vectorización. Schwartz *et al.* [2015] desarrollaron un sistema de vectorización de palabras basado en patrones simétricos. Dado un corpus C de vocabulario dimensión V y un conjunto de patrones simétricos P, el modelo obtiene una matriz simétrica M de dimensión $V \times v$ en donde $M_{i,j}$ está dado por la suma de las co-ocurrencias de las palabras w_i y w_j en los patrones P. Posteriormente, se obtiene M^* mediante la *Pointwise Mutual Information* (PMI)[4] positiva de M. El vector v_i para la palabra w_i estará dado por el renglón i de la matriz M^*. Con el fin de reducir la dispersión de la matriz, se puede aplicar un suavizado. Por cada palabra w_i se toman las n palabras con los vectores con la menor distancia coseno: W_i^n

$$v_i' = v_i + \alpha \cdot \sum_{v \in W_i^n} v$$

donde α es un factor de suavizado.

Este sistema de *embeddings* tiene la ventaja de que se puede adaptar para tomar en cuenta de manera especial los patrones que extraen antónimos (AP por las siglas en inglés de: *antonym*

[4]La PMI es un coeficiente de información mutua o asociación estadística entre dos eventos, para el cual se busca la relación entre la probabilidad de los eventos en conjunto con respecto a los eventos por separado: $PMI = log\left(\frac{p(x,y)}{(p(x)p(y)}\right)$ [Church y Hanks, 1990]

patterns). Si se toman dos matrices diferentes similares a M^*: M^{SP} y M^{AP}, donde M^{SP} toma en cuenta sólo los patrones sintácticos que no forman parte de los patrones antónimos (donde SP se refiere a los patrones simétricos que no forman parte de los patrones antónimos (AP)), y M^{AP} toma en cuenta solo los patrones antónimos, entonces es posible obtener una matriz de co-ocurrencia sensible a los antónimos:

$$M^{+AN} = M^{SP} - \beta \cdot M^{AP}$$

donde β es un parámetro de ponderación para los antónimos.

Como se puede apreciar, en general para el enriquecimiento de sistemas distribucionales, los patrones se pueden usar para estudiar y obtener relaciones horizontales, mientras que las ontologías se suelen usar como fuente de relaciones verticales. Más allá de estas dos macro categorías anteriores (palabras léxicas y funcionales), es común observar el uso de una mayor división, sobre todo para las palabras léxicas (verbos, adjetivos, nombres, etc.). Estas categorías también se pueden analizar en secuencias que dan lugar a patrones de partes de la oración o incluso a estructuras sintácticas más complejas, como árboles sintácticos [Alarcón *et al.*, 2007]. Lamentablemente, los análisis de esta naturaleza implican un texto etiquetado o bien un analizador muy confiable, una herramienta con la que se puede contar para textos bien estudiados y bien escritos, pero no para textos especializados o de escritura libre como las redes sociales.

Por otro lado, tenemos los patrones léxicos; para este tipo de patrones no es necesario tener etiquetas ni mayor análisis de información sintáctica, pero su desventaja radica principalmente en su poca flexibilidad, por lo que capturan estructuras de baja complejidad y con muy poca cobertura.

En esta investigación se procuró llegar a un punto medio entre las dos aproximanciones anteriores, ya que se hace uso de un patrón de patrones, lo que le da al sistema versatilidad y tolerancia a diferentes formas de escritura, al mismo tiempo que se mantiene sin la necesidad de análisis ni analizadores sintácticos.

Capítulo 4

Preliminares en busca de oposición contextual

Como hemos visto, el enriquecimiento automático de recursos léxicos es un problema de mucho interés, ya que sus aplicaciones son numerosas, pero su producción manual es muy costosa [Biemann *et al.*, 2018; Hearst, 1992; Snow *et al.*, 2005]. Dentro de los recursos léxicos de mayor interés se pueden encontrar, entre los más relevantes, las ontologías, las cuales ofrecen especificaciones de clases, funciones, objetos y relaciones entre el vocabulario que comparte algún dominio [Gruber, 1993]; las ontologías constituyen una herramienta muy importante para los sistemas computacionales, ya que constituyen una estructura de conocimiento externa, por lo que pueden ser reusadas a través de muchas tareas. WordNet [Miller, 1995] se muestra como un buen ejemplo de ontología y es una de las más utilizadas en la literatura.

Para expandir automáticamente estos recursos es necesario, primero, identificar los distintos tipos de relaciones paradigmáticas entre las palabras, como sinónimos y antónimos, así como su correcta clasificación. El enfoque más común para enfrentar estos problemas es mediante el uso de patrones formados manualmente, o bien, mediante bases de conocimiento y técnicas de aprendizaje supervisado.

El proceso para construir una ontología confiable de forma automática involucra la combinación de tareas independientes que cambian principalmente en función del tipo de relación.

Por lo tanto, para lidiar con tareas de esta naturaleza, el objetivo radica en la detección y extracción de relaciones particulares que se puedan encontrar en un texto crudo.

La capacidad de encontrar cómo las palabras se relacionan entre ellas siempre ha sido crucial a lo largo de la historia del procesamiento de lenguaje natural. A pesar de que los sistemas de *embeddings* distribucionales se han convertido en un punto de inflexión en el área, aún hay aspectos particularmente complicados que estos sistemas no pueden capturar, uno de ellos es la capacidad de manejar de forma diferente distintos tipos de relación. Actualmente, en muchas tareas como traducción automática, desambiguación de sentido, resumen automático, entre otras, se usan herramientas como fuente de información para suplir esta demanda de información, tal como las ontologías, relaciones como sinónimos, antónimos, hiperónimos, etc.

Como hemos visto, se pueden obtener mejoras en tareas semánticas al añadir a los sistemas de información proveniente de bases de datos de naturaleza ontológica, razón por la cual se ha buscado aprovechar la extensa disponibilidad de texto con la intención de encontrar relaciones propias de una base léxica como *WordNet* con procesos que resulten menos costosos que la recopilación manual.

En este mismo sentido, para cumplir con el propósito de la presente investigación, se desarrolla una metodología en la que se busca contribuir en el enriquecimiento de las representaciones de palabras, mediante la aportación de información adicional proveniente exclusivamente del texto; asimismo, se busca que dicha información pueda ser aprovechada en una forma similar a un recurso lingüístico y mantener, en la medida de lo posible, la independencia del idioma.

Pero para poder llegar a una metodología eficaz, en primer lugar, se realiza una exploración de las estructuras y patrones formadas por palabras vacías o funcionales. Con base en esta exploración, que en esta tesis se propone como ventanas funcionales, se diseña finalmente una metodología que sea capaz de extraer patrones de manera automática, de tal forma que se logre la extracción de pares de palabras con oposición (antónimos y co-hipónimos) a partir de texto plano.

En esta tarea se considera que palabras que contrastan o denotan oposición se llegan a usar cerca y, en esos casos, la estructura que rodea a ambas palabras es la misma. Por tanto,

los patrones sintácticos de yuxtaposición se ensamblan de tal manera que puedan extraer esos pares de palabras de una forma no supervisada. Con ello, el método logra obtener información a partir de un fenómeno de redacción que por lo regular pasa desapercibido.

Siempre que se cumpla lo anterior, será posible llevar a cabo la extracción de grupos de palabras relacionadas, y con estos grupos se puede proceder a enriquecer los *word embeddings* (en particular, se hace uso del sistema *GloVe*), lo cual nos permite evaluar la aportación lograda. De forma independiente, también se hace un etiquetado manual sobre los mismos pares de palabras extraídos para evaluar también la calidad de la relación de oposición, así como su interpretación dentro de categorías conocidas (co-hiponimos / antónimos).

4.1. Corpus

Todos los experimentos fueron ejecutados sobre los *dumps* de las Wikipedias en español y en inglés de enero del 2018.

Según su misma página[1], la Wikipedia es una enciclopedia en línea colaborativa creada y mantenida por voluntarios. Todos sus artículos están liberados bajo la licencia de *Creative Commons*, lo que la convierte en una fuente de contenido que cualquiera puede reusar y redistribuir libremente sin cargo. Más aún, Wikipedia ofrece copias de todo su contenido en forma de copias de seguridad o *dumps* que realiza de forma semanal[2]; esta característica la convierte en una fuente ideal para el análisis de información textual, por lo que se ha usado ampliamente en estudios de investigación de una gran variedad de dominios.

Español: Para los propósitos de esta investigación, se usó el *dump* de enero de 2018, el cual constituye 3.4GB de texto plano, lo que representa más de 500 millones de palabras en total y más de 10 millones de palabras únicas.

Inglés: De igual manera, se usó también el *dump* de enero de 2018, el cual constituye 8.6GB de texto plano, lo cual representa más de 1000 millones de palabras en total y casi 25

[1] `https://en.wikipedia.org/wiki/Wikipedia`
[2] `https://en.wikipedia.org/wiki/Wikipedia:Database_download`

millones de palabras únicas.

Como parte de pre-procesamiento y normalización de los datos, se utilizó el programa *WikiExtractor*[3] para la eliminación de *boilerplate*[4] y también se convirtieron todas las letras a minúsculas.

4.2. Ventanas funcionales

Una fuente de información, distinta de la co-ocurrencia, que se ha utilizado exitosamente es la distribución de frecuencia de las palabras dentro de un texto, es decir, la caracterización de palabras, patrones o estructuras específicas según qué tan común o probable es encontrarlas. Por ejemplo, Fung y Church [1994], al examinar vocabularios de documentos de dominios similares, encontraron que la frecuencia relativa de palabras como "*fisheries*" en inglés era similar a la frecuencia relativa de "*pêches*" en francés; gracias a eso, utilizaron dicha información para comparar y hacer alineación de textos paralelos.

Cuando se observa una distribución de frecuencias, se puede observar que las palabras más comunes de un texto, independientemente de cuáles sean, siempre son las palabras funcionales. Esta observación es la que inspira el planteamiento del uso de este tipo de palabras en esta investigación.

Los experimentos que se muestran en esta sección tienen como objetivo la exploración de las posibilidades que pueden tener las palabras funcionales y los contextos de las mismas para agregar información extra, además de la co-ocurrencia de palabras que se pueda encontrar dentro de los mismos textos.

De esta manera, la metodología de ventanas funcionales, aquí propuesta, consiste en la estructuración de la información del corpus en español para facilitar la búsqueda de patrones en los que participan palabras funcionales. Dado que la capacidad de estructuración de las palabras funcionales trabajan en múltiples niveles sintácticos (sintagmas, proposiciones, oraciones), fue

[3]Esta herramienta pertenece a Giuseppe Attardi y se encuentra disponible en: `https://github.com/attardi/wikiextractor`

[4]Se le llama *boilerplate* a la parte de los documentos que forman parte de la estandarización, en el caso de *Wikipedia*, esto se refiere a código dentro de los artículos que le da formato, estructura, etc.

necesario acotarse a un grupo sintáctico específico. Para este primer acercamiento, se consideró idónea la exploración de las construcciones denominadas unidades fraseológicas [Iliná, 2000; Larreta Zulategui, 2002; Ganuza, 2006], en específico las colocaciones.

En primer lugar, fue necesario elegir los tipos de palabras funcionales que servirían para nuestros experimentos, por lo que se obtuvo una lista de palabras funcionales según sus categorías: preposiciones, artículos y conjunciones [Álvaro, 1983]. Para las palabras de contenido simplemente se utilizaron todas aquellas que no formaran parte de la siguiente lista de elementos funcionales.

artículos: el, la, los, las, un, uno, una, unos, unas, lo, al, del.

preposiciones: a, ante, bajo, cabe, con, contra, de, desde, durante, en, entre, hacia, hasta, mediante, para, por, según, sin, so, sobre, tras, versus, vía.

conjunciones: a causa de que, a condición de que, a efecto de que, a fin de que, a fuerza de que, a la vez que, a medida que, a menos que, a no ser que, a pesar de que, a poco que, a punto de que, a tal punto que, a vez que, además de que, adonde, al igual que, al par que, al paso que, al punto que, al tiempo que, aun, aunque, bien entendido que, bien sabe Dios que, casi que, como, como que, como quiera que, con objeto que, con solo que, con tal de que, con tal que, conque, cuando, dado que, de ahí que, de aquí que, de modo que, de suerte que, de tal manera que, desde el momento en que, desque, donde, e, empero, en caso de que, en cuanto que, en el supuesto que, en lugar de que, en tanto que, en vez de que, en vista de que, entonces, entre que, entre tanto que, ergo, excepto que, fuera de que, gracias a que, gracias al que, hasta cuando, hasta tanto que, incluso, lo malo es que, luego, luego que, mas, merced a que, mientras, mientras que, ni, ni que, no obstante, no sea que, o, ó, o sea, ora, para que, pero, pese a que, por la cuenta que, por mucho que, por razón de que, porque, pues, pues sí que, que, respecto a que, sea, según que, si, si no es que, sin contar con que, sino, sino que, siquier, siquiera, so pena de que, so pena que, supuesto que, tal y como, tan pronto como, tanto más que, tanto menos que, u, una vez que, y, ya, ya que.

A partir de la lista de elementos funcionales (como se puede apreciar, algunos consisten en más de una palabra), se realizó una extracción de pares de palabras de contenido que se relacionaran con un elemento funcional:

$$C_1 \ F \ C_2 \tag{4-1}$$

En donde C_1 es una palabra de contenido, F es un elemento funcional, y C_2 es una segunda palabra de contenido. En la tabla 4-1 se muestran los primeros resultados más frecuentes de dicha extracción. En esta tabla, podemos ver que las palabras que se obtienen son en varias ocasiones parte de expresiones, por ejemplo "debajo del umbral" o "dio a conocer". No obstante, en la mayoría de los casos las expresiones que se obtienen tienen un gran sesgo hacia la naturaleza del corpus (enciclopédico); sin embargo, esto ya refleja relaciones, aunque sea de un modo indirecto, pues cuando se habla por ejemplo de "especie de ..." cualquier palabra que logre completar esa expresión se verá relacionada por la cualidad de ser parte de una clase.

Mas adelante, se buscaron construcciones comunes que fueran capaces de intercambiar una de sus partes a modo de encontrar más palabras compatibles con la construcción; es decir, se busca un patrón que permita el intercambio de sus componentes para lograr múltiples coincidencias relacionadas de forma similar a lo que se ve en los patrones simétricos (sección 2.4.3). Lo anterior lleva al planteamiento del experimento como una búsqueda de ventanas que incluyan dos palabras de contenido de forma similar a lo que se lleva a cabo con la expresión 4-1 (CFC), dichas palabras serán las palabras de interés para buscar relaciones; para completar el patrón, se busca que la expresión se encuentre rodeada de elementos funcionales de la siguiente manera:

$$F_1 \ C_1 \ F_2 \ C_2 \ F_3 \tag{4-2}$$

Donde las F_n representan elementos funcionales, mientras que las C_n representan palabras de contenido. Esto cumple dos funciones: por un lado, los elementos funcionales son los encargados de establecer la parte constante del patrón que se está buscando, esto es, la parte que no será intercambiable; por el otro lado, los elementos funcionales también logran delimitar las palabras

Tabla 4-1: Pares de palabras más comunes que se encuentran en el corpus alrededor de un elemento funcional

$\mathbf{C_1}$	$\mathbf{C_2}$	**F**
especie	coleóptero	de
especie	peces	de
debajo	umbral	del
especie	arácnido	de
medalla	plata	de
continuación	brinda	se
especie	pez	de
selección	fútbol	de
dio	conocer	a
pez	mar	de
género	plantas	de
especie	anfibios	de
medios	comunicación	de
especie	mantis	de
compitió	piragüismo	en
consejo	seguridad	de
especie	roedor	de
localidad	croacia	de
largo	toda	de
todos	tiempos	los

de contenido de forma efectiva, evitan en cierta medida que la estructura separe palabras que están diseñadas para estar juntas formando una sola idea (como "Estados Unidos") u otras conexiones de palabras de contenido como verbos o adjetivos, es decir, aprovechamos que las palabras vacías suelen presentar la función de conectar elementos diferentes para usarlas como punto de segmentación.

Con esta estructura, se puede entonces reemplazar una de las dos partes de los pares encontradas para encontrar palabras relacionadas e intentar formar campos semánticos alrededor de aquella palabra que se mantiene fija. En la tabla 4-2a se muestran las palabras asociadas a "coleóptero", al tomar como eje la palabra "especie" dentro de la ventana funcional: *una - de - de la*. De igual manera, en la tabla 4-2b se muestra el caso inverso, las palabras asociadas a "especie" al tomar como eje la palabra "coleóptero".

Como se ve reflejado en dichas tablas, la distribución por medio de palabras funcionales permite una asociación y relación, no sólo en cuanto a un campo semántico compartido, sino también por el uso y contexto en el que se han usado en estructuras de escritura, en una selección determinada. Se consiguen primero un grupo de especies y seres vivos; después, un modo de clasificación.

De forma similar, en la tabla 4-3a se muestran las palabras asociadas a "fútbol" al tomar como eje la palabra "selección" dentro de la ventana funcional: *la - de - de*. De igual manera, en la tabla 4-3b se muestra el caso inverso, esto es, las palabras asociadas a "selección", al tomar como eje la palabra "fútbol".

Gracias a los ejemplos de las tablas 4-3 y 4-2 es más fácil visualizar cómo se agrupan palabras mucho más relacionadas entre ellas, formando campos semánticos. Cada patrón de ventanas funcionales da lugar a dos campos semánticos: el primero, al permitir cambios en la segunda palabra (mostrado en la sub-tabla de la izquierda, en ambas tablas ejemplo); y el segundo, al permitir cambios en la primer palabra (mostrado en la sub-tabla de la derecha, en ambas tablas ejemplo).

Esto refleja la importancia que tiene el orden de las estructuras para mostrar relaciones significativas. Es decir, cuando se buscan conjuntos de palabras para palabras posteriores (sub-

Tabla 4-2: Campos semánticos de **coleóptero** y **especie** para la ventana funcional: **una especie de coleóptero de_la**

(a) Una especie de X de la

$\mathbf{C_2}$	**Frecuencia**
coleóptero	17385
peces	15140
pez	2760
anfibios	2268
mantis	2055
roedor	1380
murciélago	819
ave	728
gecos	263
musaraña	230
escarabajo	223
ránidos	161
serpientes	157
planta	156
rana	143
escincomorfos	130
anfibios	102
tortuga	96
hongo	82
lepidoptero	77

(b) Una X de coleóptero de la

$\mathbf{C_1}$	**Frecuencia**
especie	17385
subespecie	560
género	5

Tabla 4-3: Campos semánticos de **fútbol** y **selección** para la ventana funcional: **la selección de fútbol de**

(a) La selección de X de

$\mathbf{C_2}$	**Frecuencia**
fútbol	2336
balonmano	367
baloncesto	323
rugby	232
voleibol	53
béisbol	38
polo	33
críquet	16
costa	14
sóftbol	10
waterpolo	9
fútsal	7
selección	6
básquetbol	6
féminas	5
atletismo	4
natación	3
cricket	3
áreas	2
netball	2

(b) La X de fútbol de

$\mathbf{C_1}$	**Frecuencia**
selección	2336
asociación	474
federación	379
liga	193
escuela	52
confederación	46
unión	24
eurocopa	15
copa	11
sección	5
división	4
academia	4
cancha	3
mundial	3
rama	2
nacional	2

tablas izquierdas), en general, se obtienen palabras relacionadas de una forma más cercana que si se buscan en sentido contrario. Como ejemplo más enfocado, podemos analizar particularmente la tabla 4-3, donde se observa que la sub-tabla izquierda agrupa un campo semántico bien definido: deportes. En contraste, la sub-tabla derecha agrupa palabras asociadas al deporte, pero varían más al encontrarse en grupos dispersos (unos son grupos o formas de asociación, edificios o puestos).

Esto sigue cierta lógica, si se considera que el significado de una oración, en general, se va haciendo más específico conforme la oración contiene más complementos, por lo que es necesario partir de un significado más general (en este caso, un sustantivo cualquiera), a uno particular (otra serie de sustantivos que califican al previo).

4.2.1. Genitiva

La palabra genitiva es aquella que tiene una connotación de generar; para el caso del español, esta propiedad está presente en la palabra: "de". Esta preposición, al unir dos palabras, es capaz de relacionarlas de formas muy diversas; puede marcar, por ejemplo, una relación de propiedad, posesión, pertenencia, origen o materia, entre otras; es una de las preposiciones con mayor frecuencia de uso, mayor diversidad y heterogeneidad en sus funciones [Cruz Domínguez, 2011]. Esta variedad de relaciones hacen que esta palabra sea muy particular entre las otras preposiciones.

La mayoría de los patrones que se encontraron con el método de ventanas funcionales presentaron un comportamiento muy similar al de las tablas que se presentan como ejemplo en los resultados anteriores; la estructura que muestran, en relación a la preposición "de", es consistente con la diversidad que se plantea en la literatura. Los resultados parecen reflejar que, en los contextos que se extrae esta palabra en particular, se usa de una manera preferente para la unión de términos distintos en un solo concepto, por lo que su presencia se da principalmente en la posición F_2. Sin embargo, fue posible notar diferencias importantes, cuando dicha preposición se utilizaban en la posición de F_1.

En las ventanas funcionales, la preposición "de" fue, por mucho, la más utilizada para

relacionar un par de palabras de contenido; encontrar ejemplos en donde aparece "de" en la F_1 nos lleva a pensar que, en estos casos, en realidad, se están haciendo cortes "erróneos" o por lo menos de una naturaleza distinta) en construcciones que deberían mantenerse unidos, lo cual culmina en malos resultados de agrupación de palabras en campos semánticos.

La observación anterior nos lleva a plantear ciertas consideraciones sobre los resultados de un par de campos semánticos que se obtuvieron de los contextos y que se pueden usar como ejemplos: "Del consejo de X de las" y "De todos los X en". En los dos casos anteriores, tenemos la presencia de "de" (o "del") en la F_1, por lo que se realizó una extracción de palabras relacionadas (mediante la variación de la posición X), como se ha mostrado en casos anteriores; ambos campos se presentan a continuación:

Del consejo de X de las: seguridad, rectores, redacción, administración, turismo.

De todos los X en: tiempos, cargos, santos, países, mamíferos, ciudadanos, tipos, tiempo, niños, niveles, puntos, derechos, existentes, trabajadores, siglos, programas, nacimientos, laboratorios, partidos, ejércitos.

Estos son los resultados menos favorables que se obtuvieron sobre campos semánticos, pero al mismo tiempo, son los únicos que comparten una característica en común: "de" o "del" como F_1.

Si se consideran, entonces, las preposiciones "de" y "del" como un indicador de que la expresión estaba mal cortada en los contextos anteriores, se buscó un contexto mayor para los patrones anteriores. Tal búsqueda arrojó resultados como los siguientes:

- La **resolución** del consejo de seguridad
- El **líder** de todos los tiempos

Al aplicar de nuevo la metodología con estas nuevas entidades, se encontraron campos semánticos mejorados:

- resolución, reunión, presidencia, creación, decisión, aprobación, autoridad, presidenta, secretaría, sede

- líder, record, mejor, mundo, cine, ranking, cómic, segundo, metal

Estos resultados parciales parecen apoyar la idea anterior y, por consecuencia, vislumbrar que con patrones de palabras funcionales determinados puede delimitarse una amplia variedad de unidades poliléxicas, sean locuciones, colocaciones, unidades sintagmáticas verbales e, inclusive, unidades suboracionales.

Aunado al hecho anterior de no utilizar la preposición "de" fuera de la posición F_2, también se puede observar una clara tendencia a encontrar artículos en la posición F_1. Esto último ocasiona la aparición de un sustantivo como primera palabra de contenido, y el "de" obliga a que la segunda palabra de contenido también sea un sustantivo. Si bien este fenómeno puede parecer intrascendente, fue la base del diseño del experimento que se plantea con más detalle en la sección 5.1 y que se convierte en la base del mecanismo que es capaz de explotar un fenómeno al que, hasta donde se ha investigado, no se le ha puesto atención hasta ahora.

4.2.2. Ventanas funcionales en inglés

Con la finalidad tanto de verificar que el fenómeno que se observó en la sección anterior no fuera exclusivo del español, como para ampliar las posibilidades de exploración y evaluación de los resultados, los experimentos de ventanas funcionales se repitieron para el inglés. Por supuesto, se cambió la lista de elementos funcionales[5]:

articles: a, an, any, the, that, one, some, few.

prepositions: about, below, excepting, off, toward, above, beneath, for, on, under, across, beside, besides, from, onto, underneath, after, between, in, out, until, against, beyond, in front of, outside, up, along, but, inside, over, upon, among, by, in spite of, past, up to, around, concerning, instead of, regarding, with, at, despite, into, since, within, because of, down, like, through, without, before, during, near, throughout, with regard to, behind, except, of, to, with respect to.

[5]Podemos encontrar una gran variedad de palabras dentro de las categorías que se han usado, solo se tomó una muestra; se pueden encontrar más ejemplos en: `https://7esl.com/`

conjuctions: and, nor, but, or, yet, so, both, either, neither, also, but also, wheter, not only, after, although, as, as if, as long as, as much as, as soon as, as though, because, before, by the time, even if, even though, if, in order than, in case, lest, once, only if, provided that, since, so that, than, though, till, unless, until, when, whenever, where, wherever, while.

En la tabla 4-4 se muestran las relaciones más comunes de palabras alrededor de una funcional.

Tabla 4-4: Pares de palabras más comunes que se encuentran en el corpus de inglés alrededor de un elemento funcional

$\mathbf{C_1}$	$\mathbf{C_2}$	**F**
species	beetle	of
was	member	a
member	parliament	of
species	moth	of
village	municipality	and
genus	moths	of
is	member	a
species	plant	of
remainded	operation	in
has	population	a
bachelor	arts	of
secretary	states	of
species	bird	of
pursue	career	a
species	frog	of
be	part	a
coat	arms	of
chieff	staff	of
became	member	a
medal	honor	of

Al igual que como ocurre para el español, predomina la unión de sustantivos mediante

funcional genitivo. Los resultados tienen un gran parecido y se comportan de la misma manera al extraer campos semánticos mediante el intercambio de una de sus palabras. Esto nos demuestra que esta manera de relacionar palabras no es exclusiva del español, lo que nos habla de cierto grado de independencia con respecto al idioma, por lo menos para aquellos idiomas que utilicen conectores como las preposiciones.

En la tabla 4-5a se muestran los resultados de cambiar la palabra "*member*" en el patrón "*and was a member of*". Se eligió este ejemplo para mostrar cómo los campos semánticos se conservan a pesar de que en las palabras de contenido aparecen verbos. Esto también ocurre en el español, por ejemplo con el patrón "compitió en piragüismo", pero no se había mostrado aún en las tablas anteriores.

Con esto se puede observar que la metodología funciona no sólo para el español. Aún no tenemos experimentos en otros idiomas, pero como se ha mencionado antes, todo parece indicar que este mismo fenómeno se puede observar en lenguas donde existan conectores similares a las preposiciones. Es importante señalar que el cambio de la primer palabra también funciona como en los casos anteriores. En la tabla 4-5b se pueden observar las palabras relacionadas con el verbo del patrón; al buscar aquellas palabras que puedan tomar el lugar de "was" se obtienen verbos similares, verbos que caen en una misma categoría, como: "is, became, remained, were". En este caso de ejemplo se puede apreciar que se extraen palabras altamente relacionadas con el verbo en cuestión, en su mayoría, otros verbos relacionados con el significado de "ser", "convertirse en", "mantenerse siendo" o "llegar a ser". Esto es, otros verbos pseudocopulativos.

Falta considerar, sin embargo, que los verbos copulativos, y los auxiliares también, suelen ser considerados palabras funcionales, sobre todo en el inglés. En la siguiente sección mostramos experimentos que resultan al tomar en cuenta esta misma consideración.

4.2.3. Otras funcionales

En los experimentos anteriores, se consideraron únicamente palabras funcionales de categorías gramaticales claramente funcionales también. Además de artículos, preposiciones y conjunciones, es común que se tomen en cuenta pronombres como palabras funcionales. En nuestros

Tabla 4-5: Campos semánticos en inglés al utilizar verbos en posición de palabras de contenido.

(a) And was a X of

$\mathbf{C_2}$	**Frecuencia**
member	12414
part	937
director	335
fellow	267
founder	239
recipient	227
friend	226
student	220
proffesor	198
supporter	181
pupil	172
trustee	169
graduate	163
justice	118
descendant	117
resident	94
brother	88
contemporary	86
leader	81
fan	80

(b) And X a member of

$\mathbf{C_1}$	**Frecuencia**
was	12414
is	6915
became	3006
later	214
remained	132
currently	107
were	105
then	89
being	86
becoming	82
made	79
become	78
are	67
thus	62
remains	56
elected	53
becomes	43
been	39
appointed	37
not	36

primeros experimentos no los consideramos, debido a la función que pueden desempeñar como sujetos de una oración (principalmente en inglés). Los verbos auxiliares y verbos copulativos también es común verlos listados entre las palabras funcionales del inglés, no así entre las del español, en cuyo caso es poco frecuente que se tomen verbos como palabras funcionales. En cualquier caso, tampoco se consideraron para los experimentos antes mostrados.

En las tablas 4-6a y 4-6b se muestran un par de ejemplos más en los que se repite el proceso de los experimentos anteriores. Pero esta vez con el verbo más común que aparece en el corpus de inglés: *was*.

Tabla 4-6: Campos semánticos en inglés al utilizar verbos en posición de palabras funcionales.

(a) The film was X on

$\mathbf{C_2}$	**Frecuencia**
released	8461
based	1170
shot	1063
made	152
shown	123
broadcast	111
launched	64
produced	56
premiered	46
featured	39
included	33
named	31
aired	31
available	29
screened	26
put	26
announced	24
number	23
filmed	22
also	20

(b) The X was part of

$\mathbf{C_1}$	**Frecuencia**
area	9253
village	541
municipality	382
event	342
town	319
party	235
region	226
parish	222
division	189
regiment	185
city	167
song	161
tournament	139
film	132
station	106
battle	96
land	90
island	85
game	81
district	79

Dichos resultados nos han mostrado que, efectivamente, al usar ciertos verbos, se pueden conseguir relaciones de palabras y campos semánticos con un comportamiento similar al que observamos cuando se usaron palabras funcionales. Esto nos lleva a considerar a estos conjuntos de palabras también como palabras de interés para la búsqueda de relaciones léxicas. La discusión sobre la consideración lingüística de estas palabras como funcionales o no queda fuera del alcance de este trabajo.

En esencia, los experimentos de ventanas funcionales para el inglés nos demuestran la consistencia en la naturaleza de los resultados que se obtienen y son la pauta para tener una base confiable de comparación con todos los experimentos posteriores, pues estos resultados plantean la posibilidad de mantener la metodología constante, sin importar el idioma en cuestión.

Capítulo 5

Método para la extracción de pares de palabras relacionadas por oposición

Los experimentos preliminares otorgaron mucha información acerca de la manera en que se pueden aprovechar las palabras funcionales de manera que se puedan extraer palabras relacionadas. Lo que es más, el uso de palabras funcionales demostró ser capaz de obtener estructuras en que palabras de un mismo campo semántico se usan juntas, un uso que, con excepción de la genitiva (ver sección 4.2.1), es principalmente usado para denotar cierta contraposición, contraste o incluso la aclaración de características compatibles.

Para obtener los resultados anteriores, se tomaron como punto de partida una serie de palabras previamente identificadas como funcionales, lo cual demuestra la importancia que estas palabras tienen para la extracción de una estructura sintáctica de oposición. Una cualidad muy deseable en las metodologías del procesamiento de lenguaje natural es la independencia del idioma siempre que sea posible. El conocimiento previo de palabras funcionales es contrario a esta cualidad, por lo que se puede plantear una estrategia circular para la extracción, a partir de los mismos patrones, de palabras vacías de interés. Es decir, buscar desarrollar proto-patrones

que permitan la obtención de palabras clave para la mejora de los patrones nuevamente.

Todas estas observaciones dan lugar a la estructuración de una metodología de detección y extracción de pares de oposición mediante patrones sintácticos de yuxtaposición. La figura 5-1 muestra el proceso completo que se desarrolla para la extracción de pares de oposición, la cual consiste en un proceso de tres pasos consecutivos, en donde se obtienen pares de palabra que se van refinando de forma escalonada.

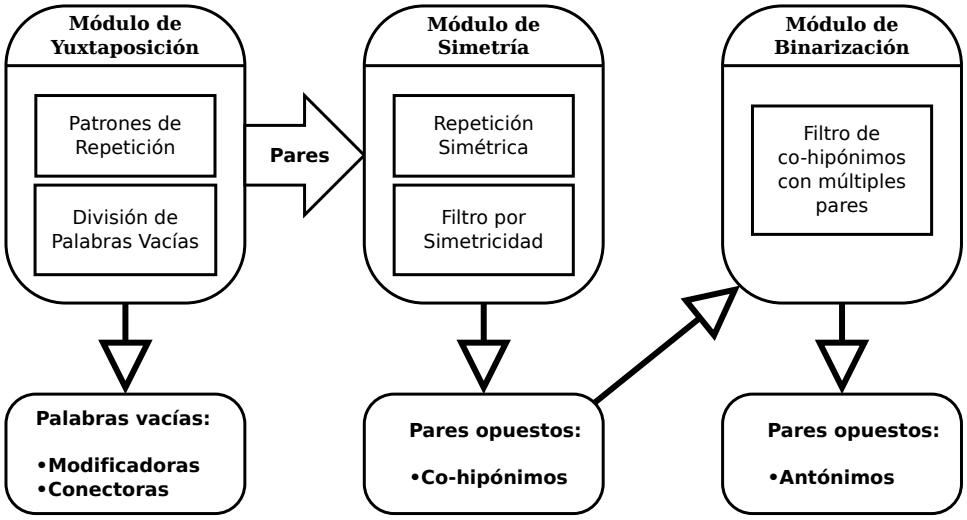

Figura 5-1: Método de extracción de pares de oposición. El proceso consiste en tres módulos, cada uno obtiene resultados que se usan como entrada del siguiente de una manera secuencial.

La yuxtaposición, el primer módulo, es la base del proceso, pues es la herramienta que permite la estructuración de un proto-patrón en el que se pueden distinguir posiciones tanto para palabras funcionales como para palabras de contenido; el proceso, por tanto, nos ofrece esas dos salidas, las palabras funcionales y los pares candidatos a tener oposición. Para filtrar los pares de entrada y lograr identificar aquellos con cierto grado de oposición se aprovecha la técnica de simetría (ver sección 2.4.3), en el segundo módulo; recordemos que la oposición es una relación simétrica, por lo que el objetivo de este segundo paso es eliminar los pares asimétricos y con eso evitar también relaciones asimétricas para lograr una calidad superior de oposición. Estos pasos son suficientes para obtener listas de palabras con un grado de oposición semejante a la co-hiponimia.

Pero el grado de oposición de los co-hipónimos, es menor al grado de oposición que se presenta en los antónimos; el tercer módulo se propone como un método para la extracción de antónimos, a partir de los pares de palabras que obtiene el módulo anterior, es decir, es el último filtro al que se someten los pares de palabras para obtener aquellos que tengan la mayor oposición y obtener antónimos a partir de un conjunto de co-hipónimos.

5.1. Módulo de repetición o yuxtaposición

Cuando se habla de repetición, no se debe entender ni como *conduplicatio* o *anadiplosis* (figuras retóricas que usan la repetición como una especie de ritmo para establecer un tema particular a nivel contextual), ni como *ploce* (figura retórica en la que una palabra se separa o repite a modo de énfasis). La repetición se debe entender más bien como aquella que uno se puede encontrar comúnmente en la estructura de la escritura natural similar al que se usa para la yuxtaposición de ideas, el tipo de repetición que no representa una figura retórica y que en trabajos como en la tesis doctoral de [Dubremetz, 2017, 11] se busca filtrar. Como Dubremetz refiere, "El lenguaje está lleno de repetición de palabras. Mientras se lee este texto, ya se encontraron con varias repeticiones sin siquiera notarlo. Y, de hecho, esto es normal; no son un evento interesante para nuestra mente".

La repetición a la que se hace referencia en Dubremetz [2017] suele pasar desapercibida principalmente por ser palabras vacías las que se repiten; además, como refiere la autora, hay bastante variedad en las razones que la ocasionan. Particularmente, la repetición más interesante para esta investigación es aquella en que las repetición de palabras vacías se da debido a la repetición de estructuras, lado a lado, una después de la otra, usadas de la misma forma. En gramática, la yuxtaposición es una de las formas de unir dos oraciones simples para formar una oración compuesta [Española y Madrid, 2001]; la diferencia con otras formas de unión es que no se usa ningún elemento conector. Lo mismo ocurre con los patrones que estamos proponiendo, ya que si bien puede haber elementos conectores dentro del patrón (como conjunciones o preposiciones), éstos también se repiten. Por lo tanto, la estructura completa (palabra funcional -

palabra contenido) queda repetida sin ningún elemento conector. Esta es la razón por la que los llamamos patrones yuxtapuestos.

Los experimentos que se han reportado hasta ahora sugieren que esa unión en general se da entre elementos que contrastan o que presentan oposición a pesar de que las razones para usar dos elementos juntos varían mucho con el contexto; sin embargo, el análisis lingüístico de la naturaleza del uso de dichas uniones queda fuera del alcance de este trabajo.

Uno de los objetivos de este trabajo es mantener al mínimo los recursos lingüísticos necesarios para la extracción de relaciones léxicas entre las palabras de un texto, ya que sólo de esta manera se pueden generalizar los procedimientos y enriquecer las representaciones de palabras de una manera totalmente no supervisada. La metodología que presentamos cumple con el objetivo, pues al basarse en un fenómeno tan frecuente y poco notorio contribuye en contar con herramientas en las que se pueda prescindir de recursos léxicos, sobre todo para situaciones en las que se trabaje con lenguas de bajos recursos o en contextos especializados. Lo anterior, aunado a que puede haber polémica en cuanto a qué se considera una palabra funcional, ha llevado al diseño del método aquí propuesto, el cual no requiere de conocer las palabras funcionales a priori.

En el método propuesto se aplica parcialmente una técnica que se suele utilizar, y que vemos en el trabajo de Schwartz *et al.* [2015], que consiste en tomar a las palabras más frecuentes como palabras funcionales. En particular, ellos toman las primeras 1000 palabras más frecuentes. Esta aproximación tiene la ventaja de ser independiente del idioma. Sin embargo, también tiene sus inconvenientes, pues dentro de las primeras 1000 palabras más frecuentes, la mayoría no son funcionales, adjetivos como *new* (rango 37), *national*(63), *american*(100) o sustantivos como *time*(52), *school*(54), *years*(57), *city*(67), *world*(68), etc.

Como se observó en la sección 4.2.1, los resultados de las ventanas funcionales nos permitieron diseñar experimentos para abordar este problema desde un nuevo enfoque al que llamamos patrones de repetición.

Por diseño, las ventanas funcionales (véase sección 4.2) nos muestran dos palabras de contenido unidas por una palabra funcional intermedia. Sin embargo, un fenómeno que se puede

observar es que en su mayoría las palabras de contenido que se unen son sustantivos, donde el primero de ellos está determinado por un artículo (el, la, los, las, etc.).

En retrospectiva, esta última observación no tiene nada nuevo, ya que dentro de la lingüística se sabe que los artículos siempre irán seguidos de un sustantivo. No obstante, para hacer uso efectivo de esa información, necesitaríamos conocer de antemano cuáles son los artículos.

Por otra parte, si combinamos ese conocimiento con la observación de que la principal forma de unir dos sustantivos es mediante otro tipo de palabra funcional (el genitivo), resulta lógico pensar que lo mismo ocurrirá para la unión de sustantivos en los que ambos cuenten con sus respectivos artículos. Y si lo anterior ocurre, entonces podemos esperar que, en algún caso, dichos artículos serán iguales.

Esto es, sin información *a priori*, buscar una palabra (candidato a artículo) junto a otra (la palabra de contenido correspondiente) que aparezca junto a ella misma (repetición del artículo) acompañada de otra (su contenido) con una palabra central separando los dos "sintagmas nominales". A continuación se presenta una expresión regular tipo *Perl* en el que se aplican las restricciones anteriores.

" (\w+)␣\w+␣\w+␣\1␣\w+"

En esencia, podemos considerar la expresión `\w+` como si fuera una palabra. La primera palabra se pone entre paréntesis para formar un grupo (el grupo 1), al cual se hace referencia mediante `\1`, por lo que en pocas palabras se están buscando cinco palabras tal que la primera de ellas se repita en la cuarta posición. Y dado que por ahora nos interesan más las palabras funcionales que las de contenido, podemos omitir la última palabra, es decir:

" (\w+)␣\w+␣(\w+)␣\1 "

En esta segunda expresión, agregamos además un segundo agrupamiento. Con esto capturamos las dos partes de la expresión que más interesan. Es importante mencionar también que a esta expresión se le agregaron posteriormente mas restricciones para evitar la repetición múltiple de una sola palabra, ya que no se tiene interés por encontrar *epizeuxis* (figura retórica en la que se repite una palabra una y otra vez para darle vehemencia o énfasis).

Hasta este momento, hemos hablado de artículos para referirnos al grupo uno, y de la funcional genitiva para referirnos al grupo dos (ya que estas son las posiciones en las que se presentan con mas frecuencia cuando se tiene una expresión como la anterior). Pero en realidad, esto sólo es un caso particular de este patrón, ya que por lo regular, en el extremo (grupo uno) quedan palabras funcionales que se utilizan para identificar cierto tipo de palabras de contenido, por ejemplo artículos para sustantivos, o ciertas preposiciones para verbos. Mientras que en el centro (grupo dos) se obtienen palabras funcionales con función de conectar sintagmas, esto incluye tanto preposiciones como conjunciones. Para facilitar la representación de estos grupos, llamaremos desde ahora al grupo uno F_1 y al grupo dos F_2 de manera similar a como se hizo en los experimentos de ventanas funcionales.

En los experimentos preliminares de la sección 4.2, una característica importante para la extracción de patrones útiles era el uso mixto de palabras funcionales (F_1 y F_2), donde el primer conjunto estaba constituido principalmente por artículos y el segundo eran sobre todo preposiciones; es decir, el primero especifica términos, mientras el segundo los conecta. Esto nos lleva a pensar que es factible dividir a las palabras vacías en dos tipos de funcionales, según el lugar en el que suelen aparecer (F_1 o F_2) y, por ende, según la función que tienen en el texto. De tal manera que si al procesar los pares extraídos se tienen palabras en F_1 que se han encontrado como F_2, quiere decir que no es necesario considerar a su F_2 como una verdadera F_2. Si se logra esta división, entonces se podrán obtener palabras relacionadas como las que se obtuvieron en los experimentos de las secciones 4.2 y 4.2.2, pero sin la necesidad de las listas de palabras funcionales previas. Gracias a esto, se puede contribuir en el propósito de mantener el método relativamente independiente del idioma y que funcione también para idiomas de bajos recursos y para contextos especializados, por ejemplo en redes sociales, donde se cambian muchas palabras y las funcionales típicas podrían no ser suficientes.

De esta manera se diseña el Algoritmo 1, en el que se arman dos listas de palabras funcionales según su posición y su confiabilidad. En la línea 7 se agrega la verificación en la que, si se encuentra una palabra que ha aparecido como F_1 en F_2, ya no se tomará en cuenta la pareja que se le haya asignado en los patrones de repetición. Además de lo anterior, se agregó otra

etapa de verificación. Ya que, conforme disminuye la frecuencia de los pares de repetición, aumenta la probabilidad de que aparezcan pares de palabras ruidosos, resultado de errores en la escritura o estructuras poco convencionales de redacción. Para evitar estos casos, se agrega al algoritmo una condición de paro, en el que si encuentra una palabra funcional típica de F_1 (y por lo tanto indicadora de que sigue una palabra de contenido) en posición de F_2, lo considera como uno de estos casos atípicos. Para definir la "palabra funcional típica" de F_1, en la línea 1 se toma la primera, que debido a que la entrada está ordenada por frecuencia, es la más común. Más adelante, en la línea 4 se establece la condición de paro, en cuyo caso, si se encuentra esta palabra como F_2, se asume que se ha llegado a un punto de resultados poco confiables y se termina el ciclo.

Algoritmo 1 Extrae pares de palabras funcionales en dos grupos F_1 y F_2

Require: Lista ordenada por frecuencia de pares de repetición ($\mathbf{L_R}$).
Ensure: Dos listas ($\mathbf{L_{F1}}$ y $\mathbf{L_{F2}}$) de palabras funcionales.

```
1:  T_F2 ← L_R[0]
2:  for all R_1, R_2 in L_R do
3:      if R_1 = T_F2 then
4:          break
5:      end if
6:      if R_2 in L_F1 then
7:          continue
8:      end if
9:      Agregar R_1 a L_F1
10:     Agregar R_2 a L_F2
11: end for
12: return L_F1, L_F2
```

5.2. Módulo de simetría

Como su nombre lo indica, en este módulo se aplica un filtro a los pares que se extraen del módulo anterior mediante el uso de simetría. Podemos recordar de la sección 2.4.3 que un patrón simétrico es aquel en el que un par de palabras X y Y pueden aparecer de manera intercambiable dentro del patrón. Ejemplos de este tipo de patrones son: "X y Y" o "tanto X como Y".

Los patrones simétricos son capaces de ser aplicados de forma independiente, sin embargo la propiedad de simetría (que las palabras pueden aparecer de manera intercambiable) se puede buscar y aprovechar en un método como el que proponemos, en donde obtenemos pares de palabras como resultado (que también provienen de patrones).

La importancia de la aplicación de simetría es clave, y está dada por una razón muy fundamental: la oposición es simétrica. Esto quiere decir que si se tiene un par de palabras opuestas, no importa en qué dirección se aplique la relación, será verdadera en ambas: "felicidad" es opuesta a "tristeza" del mismo modo que "tristeza" es opuesta a "felicidad". Esta propiedad no aplica para relaciones como la hiperonimia, por ejemplo; en esos casos la relación es asimétrica, "animal" es hiperónimo de "perro", pero "perro" no es hiperónimo de "animal".

Del módulo de repetición se extraen pares de palabras, pares que pueden estar relacionados de distintas maneras. Cada patrón da lugar a una relación potencial entre dos palabras, y si dicha relación es simétrica, se tiene la posibilidad de que ese mismo patrón será usado con las palabras invertidas. En nuestra metodología, la relación potencial está dada por la palabra funcional conectora (F_2), por lo que en el módulo de simetría se extrajeron todos aquellos pares que se encontraran en una dirección o en la otra con respecto a cada una de dichas funcionales conectoras. De esta manera, podemos excluir de los pares que se extraen todos aquellos que comprendan una relación asimétrica.

5.2.1. Simetricidad

De los trabajos que se han desarrollado sobre patrones simétricos, queda muy claro que la simetría es particularmente afín a un conjunto pequeño de palabras (como las conjunciones). El módulo de simetría es capaz de extraer una variedad de palabras funcionales conectoras, dentro de las cuales se puede pensar que el grado de simetría que mostraran será igualmente variable; esta propiedad puede ayudarnos a discriminar ciertas relaciones más útiles que otras en la búsqueda de oposición. El caso más claro de esto se podría ver en la palabra genitiva (ver sección 4.2.1), dicha palabra guarda un gran número de relaciones potenciales al ser utilizada, tantos que no puede ser confiable usarla como identificador de oposición.

Para observar de mejor manera la simetría asociada a una palabra, calculamos una puntuación de simetricidad (S) como la razón:

$$S = \frac{f_s}{f_t} \times 100$$

donde f_s es la frecuencia de intersección simétrica y f_t la la frecuencia total de repetición.

Este factor es ideal para hacer un filtrado de palabras funcionales para no tomar en cuenta aquellas palabras que puedan representar una gran variedad de relaciones, es decir, no tomar en cuenta las genitivas. El rango (posición que toma al ser ordenado de mayor a menor con respecto a una variable) de la frecuencia de una palabra puede ser muy diferente a su rango de simetricidad, por lo que el método de filtrado que se propone es no tomar en cuenta los pares que provengan de la palabra funcional con mayor caída de rango de frecuencia a simetricidad.

5.3. Módulo de binarización

Un método que ha dado buenos resultados en el enriquecimiento de *word embeddings* es el uso de información de antonimia. Mrkšić *et al.* [2016] en su investigación usaron el método de *counter-fitting* para separar las palabras antónimas entre ellas (más detalles en la sección 7). Sin embargo, algo interesante que se puede mencionar es que dentro de los antónimos que se usaron, muchas palabras en realidad pertenecían a diferentes instancias de una misma categoría (como los meses del año, o diferentes países). Esta relación entre antónimos y co-hipónimos se puede observar en muchos recursos léxicos, como ejemplo de esto se muestra la tabla 5-1 con pares extraídos del corpus de paráfrasis PPDB [Pavlick *et al.*, 2015].

En este sentido, es muy común encontrar en la práctica generalizaciones de antónimos como todos los elementos excluyentes que forman parte de un mismo campo semántico. Palabras como “carro”, “camión” y “motocicleta” también son muchas veces considerados antónimos en este tipo de recursos léxicos [Gagné y L’Homme, 2016].

Como se desarrolló en la sección 2.4.2, la co-hiponimia y la antonimia están cercanamen-

Tabla 5-1: Pares de “exclusión” extraídos del corpus de paráfrasis PPDB

afghanistan	pakistan
africa	asia
argentina	brazil
argentina	chile
argentina	germany
argentina	mexico
bad	best
bad	better
bad	good
blue	green
blue	red
blue	white
blue	yellow
children	mothers
children	parent
children	parents

te relacionadas, por lo que los antónimos se pueden considerar como un subconjunto de los co-hipónimos; la principal diferencia que permite diferenciar ambos grupos de palabras es la propiedad binaria de los antónimos. La antonimia se da entre categorías que solo cuentan con dos elementos, o bien una línea gradual de elementos con dos extremos bien definidos (por ejemplo, blanco y negro). No obstante, la discusión lingüística de esta hipótesis está más allá del alcance de este trabajo. Nosotros vamos a partir del supuesto de que si se tiene un par de palabras antónimas, también se puede establecer que esas mismas dos palabras serán co-hipónimos binarios. Esta afirmación nos permite proponer un método para extraer antónimos a partir de co-hipónimos con un algoritmo muy sencillo: A partir del conjunto de pares co-hipónimos, se extraen todos aquellos pares que contengan una palabra que no tenga ningún otro par, de tal manera que solo se obtengan aquellas palabras que tuvieron una y solo una palabra opuesta y, por lo tanto, sean mas probables de ser antónimas.

Capítulo 6

Experimentación

En este capítulo se muestran los experimentos y los resultados que se obtienen al aplicar cada uno de los módulos de la metodología propuesta sobre el corpus de Wikipedia del 2018, descrita en 4.1.

En el capítulo 5 se plantea el proceso completo que sigue la metodología para la detección y extracción de oposición a partir de un texto. Podemos recordar los tres módulos que conlleva el proceso:

Yuxtaposición: En este módulo se busca detectar patrones comúnmente desapercibidos y de donde se extraerán posteriormente los pares de palabras. De este módulo se espera detectar las palabras conectoras y modificadoras (palabras funcionales), así como las ventanas con los pares de palabras posiblemente opuestas que pasarán al siguiente módulo.

Simetría: En este módulo se aplican dos filtros a los pares de palabras, uno para detectar únicamente relaciones simétricas, y el otro para eliminar conectores ruidosos, particularmente el genitivo. De este módulo ya se espera obtener los pares de palabras con un cierto grado de oposición, palabras candidatas a ser co-hipónimas.

Binarización: El último módulo busca extraer los pares de mayor oposición entre los obtenidos del módulo anterior mediante el requisito de exclusividad de opuesto. De este módulo se esperan obtener los pares de palabras de mayor oposición, los candidatos a antónimos.

En el resto del capítulo, se muestran con detalle los resultados que se obtienen de cada uno de los módulos mencionados.

6.1. Módulo de yuxtaposición

Para comenzar el proceso, se hace uso de la expresión regular que se define en la metodología en la sección 5.1. Recordemos que dicha expresión toma la forma:

" (\w+)␣\w+␣(\w+)␣\1 "

La expresión anterior da como resultado pares de palabras como las que se muestran en la tabla 6-1 donde, si se analizan los pares de palabras que se extraen, se puede observar que la mayoría constan de palabras funcionales como se predijo que sucedería. Gracias a los resultados de los experimentos pasados, tenemos en posición conectora (F_2) palabras como "de, y, en, a, con, por", mientras que en la posición F_1 se presentan mas bien artículos "la, el, los, las". La funcional genitiva ("de") aparece más como F_2 con una gran diferencia. No obstante, también se presentan casos que difieren de lo esperado (“ciudad de”, “provincia de”, etc.) ya que no son palabras funcionales conectoras, que era lo que habíamos encontrado en los experimentos de las ventanas funcionales. Se puede suponer que la razón de que esto ocurra es debido a la repetición de la palabra funcional que acompaña a la palabra de contenido. Podemos ver un ejemplo de lo anterior en la tabla 6-2.

Como se puede apreciar, por la misma naturaleza de la posibilidad de repetición de palabras vacías, una de éstas puede tomar el lugar de una palabra de contenido (en el segundo renglón de la tabla "la" está en la posición C_1), lo que provoca que sea la palabra de contenido la que tome el lugar de la funcional o vacía ("ciudad" pasa a tomar la posición F_2). Sin embargo, cuando ocurre esto, la palabra que toma el lugar de F_1, en su mayoría, es una palabra que toma el lugar F_2 (ya sean preposiciones en general o la funcional genitiva en particular, en el ejemplo que se muestra es "de", palabra que como hemos visto antes tiene su mayor número de apariciones en la posición F_2).

Como resultado del algoritmo 1 aplicado al corpus de español, se obtuvieron los siguientes

Tabla 6-1: Pares de repetición extraídos directamente mediante expresión regular

$\mathbf{F_2}$	$\mathbf{F_1}$
de	la
y	la
y	de
ciudad	de
y	el
en	la
de	los
provincia	de
universidad	de
era	de
en	el
a	la
de	las
década	de
y	los
y	en
familia	de
con	la
por	la
serie	de

Tabla 6-2: Comparación de patrón de repetición. En primer lugar se tiene un caso esperado, en segundo lugar se tiene un ejemplo de extracción de palabra de contenido en lugar de una vacía.

$\mathbf{F_1}$	$\mathbf{C_1}$	$\mathbf{F_2}$	$\mathbf{F_1}$
la	ciudad	de	la
de	la	ciudad	de

resultados:

L_{F1}: la, el, los, las, del, su, un, al, se, the, san, más, una, lo, sus, dos, le.

L_{F2}: de, y, en, a, por, es, que, con, como, para, desde, era, o, hacia, hasta, fue, contra, durante, entre, sobre, son.

Estos resultados nos muestran que efectivamente las palabras vacías se pueden dividir, mediante su posición, en grupos que mantienen funciones distintas. Gracias a esto se pueden usar los elementos de la lista L_{F1} como artículos (o palabras con usos afines); de manera similar, los elementos de la lista L_{F2} pueden ser usados como preposiciones y conjunciones. Estos elementos, directamente extraídos del corpus, es lo que hace justamente que los métodos descritos en las secciones 4.2 y 4.2.2 puedan ser independientes del idioma.

Como un resultado extra, además de las listas de palabras funcionales se obtienen también pares de palabras de contenido. En la tabla 6-3 se observan el tipo de palabras pares que se obtienen mediante este proceso.

En la tabla 6-4 se muestran los pares que se obtienen para "fútbol", una de las palabras que se analizaron previamente en el experimento de ventanas funcionales (ver tabla 4-3a). En dicha tabla se muestran dos subtablas que denotan una diferencia fundamental de este nuevo experimento: para una palabra dada, podemos encontrar palabras relacionadas en dos posiciones distintas del patrón. En la subtabla izquierda (tabla 6-4) se muestran los pares con los que aparece la palabra "fútbol" cuando ésta aparece en posición de C_1, es decir, extraemos C_2. Mientras que en la subtabla derecha (tabla 6-4) se muestran los pares cuando "fútbol" aparece como C_2, por lo que se extraen las palabras en posición C_1.

Al analizar las tablas anteriores, se puede notar una gran similitud con los resultados que se obtuvieron de la misma palabra en los experimentos de ventanas funcionales, a pesar de tratarse de un tipo de contexto diferente ya que, como se recordará, en el experimento de las ventanas funcionales se obtenían campos semánticos a partir de una palabra pivote, lo que implicaba un cambio de contexto, mientras que para este nuevo experimento, no se hace cambio de contexto, sino que las palabras aparecen en el mismo contexto que la palabra de interés.

Tabla 6-3: Pares de palabras de contenido resultado de los patrones de repetición.

$\mathbf{C_1}$	$\mathbf{C_2}$
guerra	independencia
final	copa
presidencia	república
bandera	ceremonia
humanidad	unesco
imagen	virgen
copa	uefa
mayoría	población
capital	provincia
letra	canción
norte	sur
agricultura	ganadería
organización	federación
clasificación	copa
casa	cultura
copa	liga
santos	últimos
población	ciudad
pedro	pablo
universidad	república

Tabla 6-4: Pares de “fútbol” resultado de los patrones de repetición.

$\mathbf{C_1}$ = "fútbol"	$\mathbf{C_2}$ = "fútbol"
deporte	baloncesto
baloncesto	rugby
club	club
rugby	béisbol
país	profesionalismo
atletismo	interés
equipo	país
béisbol	tenis
tenis	atletismo
ciclismo	equipo
hockey	ciclismo
voleibol	campeonato
principal	básquetbol
año	municipio
amor	golf
salvador	debut
estado	cricket
cricket	boxeo
boxeo	balonmano
básquetbol	récord

Lo anterior implica una ventaja para los pares de repetición con respecto a las ventanas funcionales, pues no necesitan de una palabra pivote para extraer campos semánticos, sólo es necesaria la palabra de interés. Como contraparte, tenemos que la palabra pivote puede ser un recurso de desambiguación de campos semánticos al establecer un sentido para las palabras que relaciona. Por otra parte, se presenta también la desventaja de la frecuencia, en esta nueva forma de extraer campos semánticos las palabras tienen una menor frecuencia de aparición. Lo que nos lleva a pensar que se podrá alcanzar una menor cobertura, ya que son menos las palabras que se pueden relacionar de esta manera. Es por eso que para experimentos en trabajo futuro esperamos buscar la manera de combinar los patrones de repetición con las ventanas funcionales para obtener mejoras en nuestros resultados.

6.1.1. Repetición en inglés

Uno de los principales objetivos de este trabajo es que las estructuras que se analizan sean lo más independiente del idioma posible. Por lo tanto, los experimentos anteriores se repitieron para el inglés. Al repetir el experimento de los patrones de repetición, sin cambios, se obtienen también dos listas de palabras funcionales:

L_{F1}: the, a, he, his, be, she, one, two, her, new, their, its, st, an, best, they.

L_{F2}: of, in, and, for, to, on, at, from, is, with, that, was, as, or, about, by.

Y al igual que ocurre para el español, en el inglés también se presenta el equivalente de la funcional genitiva como la funcional más común de las F_2.

De la misma manera, se pudieron obtener grupos semánticos para las palabras. En la tabla 6-5 se muestran los pares que se obtienen para la palabra “sea” (mar). En dicha tabla se muestran nuevamente dos sub-tablas, una para cada posición del patrón que ocupa la palabra para extraer su campo semántico.

En dicha tabla se pueden notar una gran similitud con los resultados que se obtuvieron tanto en español como para inglés sin la necesidad de cambiar en nada el algoritmo. Estos resultados son muy importantes, ya que demuestran cierto grado de independencia del idioma para el

Tabla 6-5: Pares de “sea” resultado de los patrones de repetición.

$\mathbf{C_1}$ = "sea"	$\mathbf{C_2}$ = "sea"
east	law
north	bottom
west	surface
south	race
land	god
hebrides	edge
mountains	coast
carolinas	depths
second	shores
sky	heart
river	exploration
area	mystery
sun	land
coast	mountains
lake	waves
mountain	song
island	river
town	goddess
mouth	temperature

método. Por supuesto, se debe hacer notar las cualidades que ambos idiomas comparten: por un lado, tenemos la segmentación de palabras por espacios, nuestra metodología aprovecha esta propiedad, ya que se basa en una expresión regular que usa espacios como separador de palabras, por lo que en idiomas aglutinantes se debe tener el cuidado de llevar a cabo previamente la *tokenización* adecuada; el otro aspecto que comparten es la presencia de preposiciones, las cuales son identificadas y clasificadas por el método. En todos los experimentos que se mostraron, se asume que las palabras funcionales se posicionan antes que las palabras de contenido; en un idioma donde prevalezcan las postposiciones, ese orden se tendrá que invertir. En general, los resultados hasta este punto son buenos indicadores de que la metodología funcionará en idiomas que se puedan *tokenizar* y usen pre o postposiciones.

6.2. Módulo de simetría

Un detalle muy importante que se puede observar de la tabla 6-4 es la aparición de elementos repetidos en ambas subtablas. Como es de esperar, las palabras repetidas muestran una relación más fuerte con la palabra de interés. Pero más allá de eso, muestran una tendencia por coincidir en relación de cohiponimia con dicha palabra. Es decir, podemos ver cómo la palabra "deporte", a pesar de su gran relación con la palabra "fútbol" (hiperonimia), desaparece en la segunda tabla mientras que casi todos los otros deportes forman parte de la intersección. Una explicación que se le puede dar a esto es que la hiperonimia no es una relación simétrica.

No obstante, también se pueden encontrar dentro de las tablas varias palabras con relaciones asimétricas (país, club, equipo) que se mantienen también dentro de la intersección. En este punto fue aplicado el módulo de simetría para extraer los pares que aportan mayor información de oposición. En la tabla 6-6 se muestran los resultados de la aplicación de simetría para el caso de "fútbol", así como una muestra aleatoria de 20 elementos extra.

En los resultados, gracias a la adición de simetría, se puede observar una clara preferencia por extracción de pares de palabras que pertenecen a una misma categoría semántica, o bien, que comparten hiperónimo. Esto a su vez implica una mayor tendencia hacia la obtención de

Tabla 6-6: Intersección simétrica de pares de repetición.

(a) Pares de fútbol

boxeo
ciclismo
béisbol
hockey
tenis
club
básquet
atletismo
básquetbol
equipo
balonmano
rugby
baloncesto
cricket
golf

(b) Muestra aleatoria

música	televisión
biblioteca	oficina
ciudad	posguerra
cuadrado	triángulo
cola	garganta
inglés	italiano
escultura	música
salvación	redención
gato	ratón
plomo	cinc
descenso	ascenso
justicia	caridad
guitarrista	bajista
occidente	norte
astronomía	geografía
pino	ciprés
capuleto	montesco
arpa	piano
catalán	inglés
cómico	trágico

co-hipónimos, los cuales, como se explica a detalle en la sección 2.4.2, están muy ligados con la oposición semántica, eso sin mencionar que también es una relación léxica muy importante para las ontologías. En los ejemplos que se muestran a estas alturas ya se puede ver una clara mejora en la calidad de relación que guardan las palabras hacia la ruta que se ha marcado en los objetivos de esta investigación.

6.2.1. Filtro por simetricidad

Siguiendo nuestra metodología, en este punto se aplica el filtro por simetricidad, el cual consta de la detección del mayor cambio de rango entre las palabras funcionales en busca de la genitiva de forma automática. Es decir, como se muestra en la figura 6-1, cuando tenemos una distribución de frecuencias de las F_2 que se presentan en los pares de repetición, podemos observar a la funcional genitiva en la primera posición.

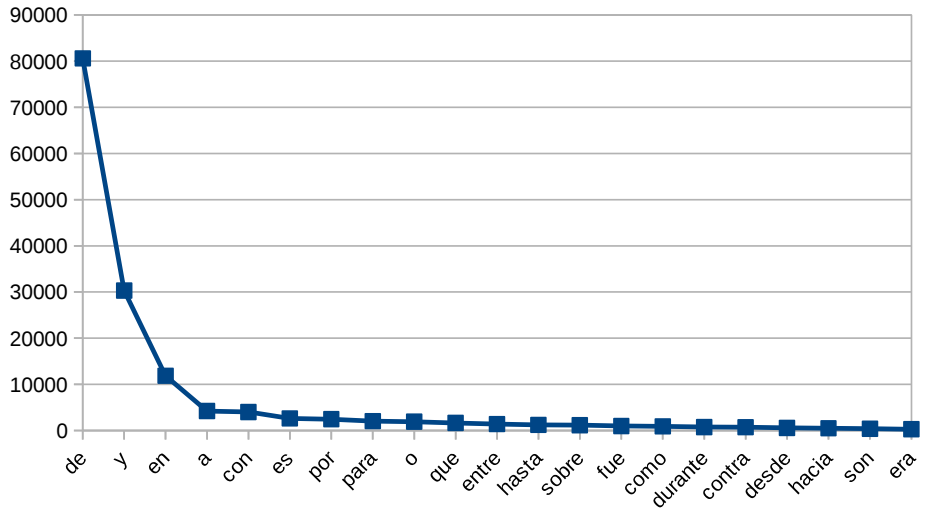

Figura 6-1: Distribución de frecuencia total de F_2

Sin embargo, como se sugería desde los experimentos preliminares (sección 4.2) la funcional genitiva establece, en su mayoría, relaciones asimétricas entre palabras, por lo que al hacer la extracción de la intersección simétrica de pares, y obtener nuevamente una distribución de frecuencia de las F_2, podemos ver una importante diferencia de la proporción de aparición de esta palabra en particular. En la figura 6-2 se muestran la distribución de frecuencias de

aparición de las F_2 después de hacer la intersección simétrica.

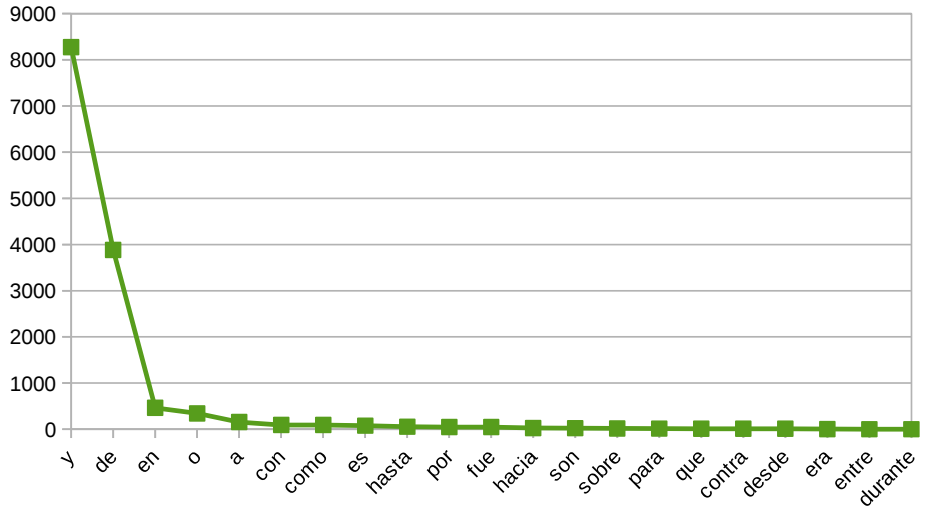

Figura 6-2: Distribución de frecuencia simétrica de F_2

A pesar de que las gráficas muestran una distribución similar, los rangos de las palabras son diferentes, lo cual implica un cambio drástico en la presencia de simetría para ciertas palabras, como el "y". Ya que, a pesar de tener una frecuencia total menor a la mitad de la frecuencia para la palabra "de", termina con más de lo doble que ésta.

En la figura 6-3 se muestra nuevamente la distribución de frecuencia de las F_2 junto con su puntuación de simetricidad. De igual manera, en la figura 6-4 se muestran los mismos datos, pero en esta ocasión se presentan ordenados por simetricidad.

Ambas gráficas reflejan información interesante. En la figura 6-3 podemos ver cómo se identifican claramente las conjunciones 'y' y 'o' para tomar las primeras posiciones, de igual manera se pueden ver identificados conectores tales como son "como", "hasta" y "hacia", que son utilizados comúnmente para agrupar contrastes y para indicar dirección entre dos lugares distintos. En realidad, el único problema que encontramos con la detección de palabras que denotan simetría fue con "son". Esta palabra es detectada por el uso que se le da para la introducción de enumeraciones o listados, por ejemplo:

- las siguientes **son** las carreras ...

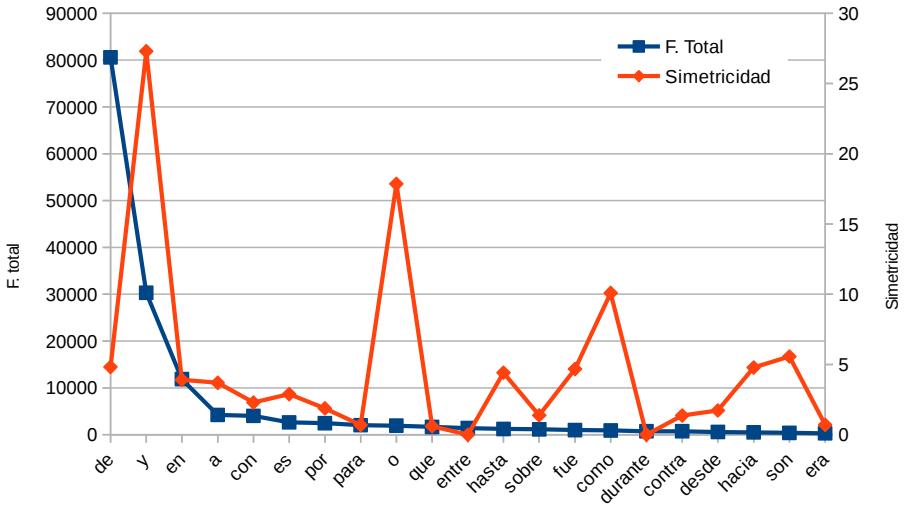

Figura 6-3: Distribución de frecuencia total y simetricidad

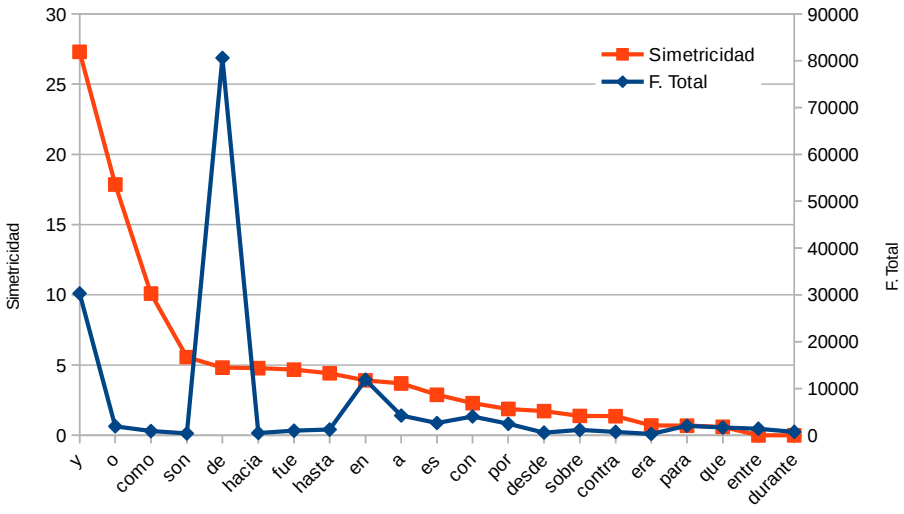

Figura 6-4: Simetricidad y frecuencias de las funcionales F_2

- ... los datos **son** los siguientes:

Por su parte, en la figura 6-3 se puede apreciar con mucha claridad la detección de la palabra funcional genitiva, lo cual es muy importante para el filtrado de pares de palabras relevantes. Hemos comentado que la funcional genitiva establece relaciones asimétricas entre palabras. Sin embargo, se usa de tantas maneras y en tantos contextos (por ejemplo "la mayoría de la acción" y "la acción de la mayoría") que tiene mucha aparición inclusive en la simetría. No obstante, sus pares de palabras simétricos no mantienen las mismas propiedades que en otras palabras de simetricidad semejante como se puede ver en el ejemplo, esta palabra combina tantos tipos de relaciones que la "simetría" en realidad solo es una relación distinta.

6.2.2. Simetría en inglés

El análisis de la simetría se inició a partir de la observación de elementos repetidos en las sub-tablas de la tabla 6-4; sin embargo, ese mismo fenómeno está presente también en la tabla 6-5, es decir, en inglés. Si se busca simetría en los campos semánticos del inglés, también obtenemos conjuntos mucho más relacionados que sin la toma de simetría. La tabla 6-7 muestra los resultados de tomar en cuenta la simetría en los conjuntos que se obtuvieron para la palabra "sea" junto con una muestra aleatoria de pares de palabras.

Los resultados muestran exactamente la misma tendencia que sus homólogos en español: una clara preferencia por extracción de pares de palabras que pertenecen a una misma categoría semántica, o bien, que comparten hiperónimo.

Por su parte, en inglés también es posible medir la simetricidad y observar el cambio de rango que se produce al comprar la frecuencia de las palabras funcionales conectoras contra la simetricidad que presentan. La figura 6-5 muestra una comparativa entre la distribución de frecuencias y la simetricidad de las funcionales F_2, es decir, la contraparte de lo que se observó en la figura 6-4, solo que ahora para el inglés.

El comportamiento es casi idéntico a lo que se puede observar en las gráficas del español: se tiene un pico que sobresale muchísimo de los valores del rededor, y dicho pico es precisamente la palabra genitiva. Por supuesto, en el caso del inglés dicha palabra es "of", pero cumple,

Tabla 6-7: Intersección simétrica de pares de repetición.

(a) Pares de "sea"	(b) Muestra aleatoria	
crew	site	museum
hills	kinks	yardbirds
lake	document	name
stars	church	patriarchate
islands	book	violence
east	biggest	state
mirror	streets	area
background	brahmaputra	ganges
rest	roof	chancel
middle	period	month
age	concept	name
beach	grounds	country
whole	finals	qualifying
port	reference	job
area	defendant	court
plain	cities	region
castle	people	mountain
cliffs	script	idea
water	front	seating
coasts	night	match

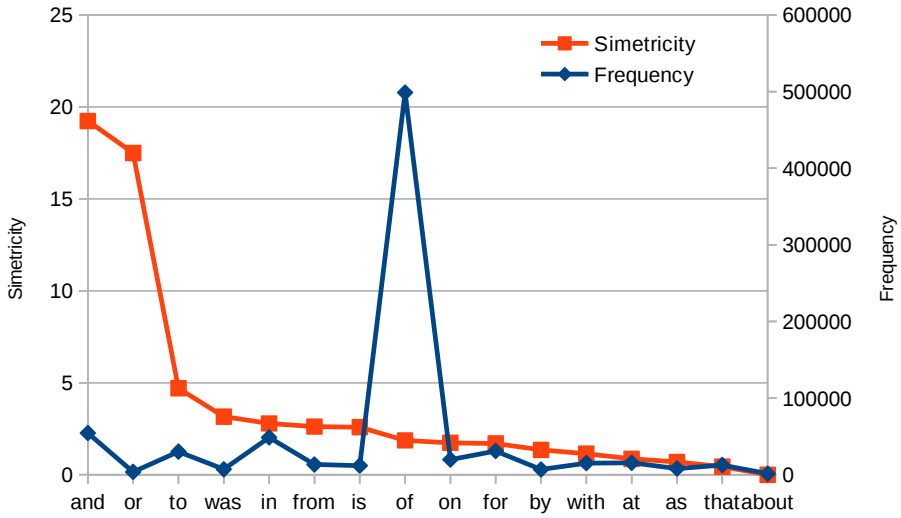

Figura 6-5: Simetricidad y frecuencias de las funcionales F_2 para el inglés

igualmente, con una variedad de funciones que, como se ha mencionado antes, desemboca en que las relaciones denotadas con esa palabra sean de menor calidad, por lo que se pueden desechar en la extracción. Nuevamente, el experimento demuestra la independencia que tiene el método y, en este caso, también el fenómeno que se está explotando, con respecto al idioma.

6.3. Módulo de binarización

Con la adición de simetría, los pares de palabras presentan una mejora cualitativa en cuanto a su capacidad de diferenciación entre elementos de una misma categoría. La categoría que mejor se encontró que describe a estos pares es la relación de cohiponimia, y en algunos casos se ve una tendencia hacia la antonimia.

Para completar la metodología y buscar extraer palabras con el mayor grado de oposición, y por tanto más probables de ser antónimos, se aplica finalmente el módulo de binarización. En la tabla 6-8 se puede observar una muestra aleatoria de 20 pares que se seleccionaron con esta restricción.

Con estos últimos resultados se obtienen pares con un gran nivel de oposición y muchos de ellos comúnmente usados como antónimos prototípicos (calor - frio , emoción - razón). Este

Tabla 6-8: Muestra aleatoria de pares binarios

audición	visión
aimara	quechua
salarios	precios
calor	frio
compradores	vendedores
consciente	inconsciente
terroristas	rehenes
docentes	alumnos
oveja	cabra
suicidio	asesinato
tejidos	órganos
diablo	ángel
semitono	tono
emoción	razón
objetos	personajes
racionalismo	empirismo
muertos	vivos
melodías	letras
ventana	puerta
aritmética	geometría

constituye el último paso de nuestra metodología para la extracción de palabras de oposición, con esto se cumple el principal objetivo, ya que ahora se cuenta con dos conjuntos de palabras: antónimos y co-hipónimos, ambos extraídos de forma no supervisada a partir de un texto plano.

6.3.1. Binarización en inglés

Más relevantes incluso para la etapa de evaluación, son los resultados de los experimentos que se logran con el mismo proceso de binariazación para el inglés. El proceso fue exactamente el mismo: a partir del conjunto de pares co-hipónimos, se extraen todos aquellos pares que contengan una palabra que no tenga ningún otro par, de tal manera que solo se obtengan aquellas palabras que tuvieron una y solo una palabra opuesta y, por lo tanto, sean más probables de ser antónimas. En la tabla 6-9 se puede observar una muestra aleatoria de pares que se obtienen mediante este método.

Exactamente igual que ocurrió para el español, las palabras ahora presentan un gran grado de oposición y muchas de ellas son antónimos prototípicos (slow - fast, beggining - end, differences -similarities). Esto quiere decir que todos los pasos de la extracción de oposición se aplican igualmente tanto para el español como para el inglés sin tener que hacer ninguna modificación en el método.

Tabla 6-9: Muestra aleatoria de pares binarios para el inglés

roots	leaves
cree	blackfoot
polymer	solvent
gay	lesbian
powhatan	english
juniors	seniors
differences	similarities
defence	prosecution
slow	fast
profane	sacred
beggining	end
lydians	medes
supply	demand
receiver	transmitter
institutions	citizens
rotating	stationary
perimeter	area
bulls	blackhawks
celestial	terrestrial
row	column

Capítulo 7

Evaluación

La evaluación de la extracción de relaciones entre palabras es una tarea compleja en sí misma. Una intuición inicial puede llevar a pensar que es posible hacer una evaluación mediante la comparación de las relaciones con recursos creados a mano; sin embargo, no se puede asegurar que el tipo particular de relación, la cobertura o el dominio de los recursos manuales sean apropiados para este tipo de comparación. Tal como mencionan Brewster *et al.* [2004], "no hay un conocimiento claro que deba ser adquirido", existen muchas formas en que las palabras se relacionan según su contexto y el dominio en el que se usan, así como la naturaleza de los textos, pero eso no implica en ningún momento que las relaciones que mantienen no tengan significado ni justificación.

Más aún, si los resultados se evalúan al compararlos con las relaciones que se encuentran en otro recurso, por ejemplo *WordNet*, entonces solo se estaría evaluando qué tan bueno es el sistema para obtener relaciones que ya se pueden encontrar en otro medio, cuando en realidad, lo ideal sería que el algoritmo obtenga muchas relaciones correctas que no se hallen en ontologías. La figura 7-1 ejemplifica tres casos de manera gráfica. El círculo azul representa el total de relaciones correctas entre palabras, el círculo rojo representa el conjunto de relaciones capturadas en un recurso manual y el círculo verde representa el conjunto de relaciones extraídas por un sistema. En primer lugar, del lado izquierdo tenemos el caso de un sistema que cumpliría perfecto con una evaluación de comparación, todas sus relaciones son correctas, pero no aporta

nada nuevo con respecto al recurso manual. En segundo lugar, en el centro, tenemos un caso de un sistema automático en el que todas las relaciones son correctas, pero distintas a las que se pueden encontrar en un recurso manual. Si bien aportaría más información, irónicamente este sistema tendría un desempeño nulo con una evaluación por comparación. Finalmente, en el lado derecho, tenemos el caso que seguramente será el más común, un sistema automático donde solo se tendrá una intersección parcial con los recursos manuales, se obtendrán nuevas relaciones que no se encuentran en el recurso manual y, como cualquier sistema automático, también tendrá relaciones erróneas.

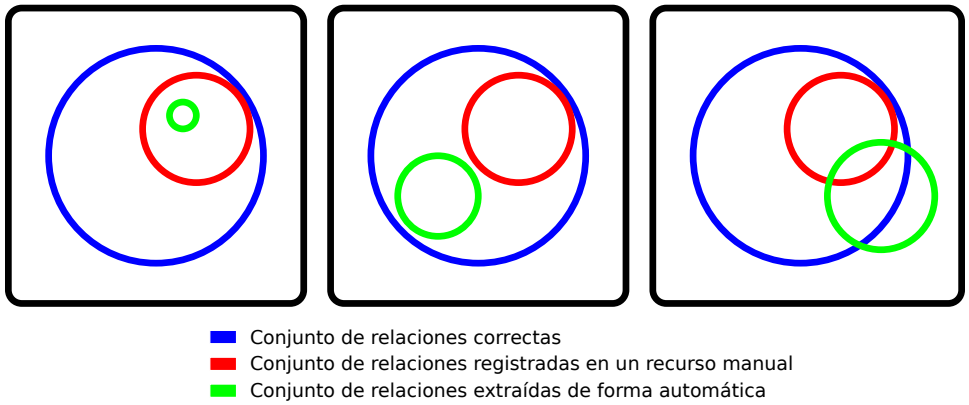

Figura 7-1: Diagramas de Venn ejemplificando tres tipos de evaluaciones de relaciones entre palabras: Evaluación perfecta sin información adicional, evaluación nula con aportación perfecta y evaluación mala con buena aportación.

Esta es la razón por la cual no es conveniente llevar a cabo una evaluación mediante la comparación con recursos léxicos existentes. En su lugar, se considera que lo más apropiado para un problema de esta naturaleza es hacer una evaluación mediante un etiquetado manual sobre las relaciones extraídas.

7.1. Co-hipónimos

Para evaluar la calidad de los pares que extrae el sistema como candidatos a co-hipónimos, se solicitó el apoyo de un par de etiquetadores ajenos a esta investigación para llevar a cabo una clasificación binaria sobre dos corpus tomando como base el trabajo de Weeds *et al.*

[2014]. En su artículo, Weeds desarrolla un sistema automático para distinguir entre relaciones diferentes que pueden pasar como "similitud" para un sistema distribucional (sinónimos, antónimos, hipónimos, hiperónimos, co-hipónimos, etc.). En particular, dos corpus etiquetados para la distinción de dos relaciones en particular: hiperónimos y co-hipónimos.

***BLESS*:** Es una colección de ejemplos de muchos tipos de relaciones léxicas (hiperónimos, merónimos, co-hipónimos, etc.) para un conjunto de 200 palabras que tienen un único significado concreto [Baroni y Lenci, 2011]. A partir de esa colección, en Weeds *et al.* [2014] se obtuvo un corpus con 5835 pares de palabras.

***WordNet*:** Una base de datos léxica que agrupa palabras en conjuntos de sinónimos, además de que conecta dichos grupos según diferentes relaciones léxicas [Miller, 1995]. A partir de esta base de datos, en Weeds *et al.* [2014] se obtuvo un corpus con 3771 pares de palabras. A diferencia del corpus que se extrae a partir de BLESS, en este caso los elementos negativos se extraen de casos de implicación (hiperónimos).

Los corpus antes mencionados tienen importantes cualidades que lo hacen ideal para la tarea que compete a esta investigación: Se trata de dos corpus que han sido cuidadosamente balanceados de tal manera que la mitad de los elementos son co-hipónimos y los elementos que no lo son, aunque son palabras que mantienen otro tipo de relación, siguen siendo palabras similares entre ellas, esto garantiza que la tarea de distinguir los co-hipónimos dentro de los corpus no se puede conseguir mediante la detección de similitud. Lograr esta tarea, por tanto, supondrá beneficios sobre solo usar sistemas distribucionales.

Para ambos casos, se usaron todos los pares distintos a co-hiponímia como elementos negativos, con lo cual se dio lugar a la conformación de dos corpus de evaluación de co-hiponimia mediante la combinación de dichos elementos negativos y los co-hipónimos que se extrajeron usando el método descrito en la sección 5.2. Estos dos corpus de evaluación fueron proporcionados a los etiquetadores para una clasificación binaria: co-hipónimo *vs* no-cohipónimo. Los resultados del etiquetado permite analizar la percepción humana sobre los pares de palabras extraídos automáticamente.

Para el caso del corpus que se construyó a partir de los datos de *BLESS*, se etiquetaron 1000 pares de palabras, los etiquetadores presentaron un porcentaje de acuerdo del 82.93 % y un coeficiente κ de Cohen de 0.6295; lo que representa un acuerdo sustancial. Para el caso del corpus construido a partir de *WordNet*, se ejecutó el etiquetado de 3000 pares. Para este caso, el porcentaje de acuerdo entre los etiquetadores fue de 62.17 % con un coeficiente κ de 0.2674. Se puede afirmar entonces que las relaciones en el corpus de *WordNet* son mas confusas que aquellas de *BLESS*; no obstante, sigue siendo un acuerdo aceptable.

La exactitud que se logra para el corpus basado en *BLESS* alcanza el 80.00 % para uno de los etiquetadores, mientras que se obtiene 82.24 % para el otro, por lo que se obtiene un promedio de 81.12 % de exactitud. Por el otro lado, en el caso del corpus basado en *WordNet*, se obtuvo una exactitud del 71.26 % para uno de los etiquetadores, y un 59.21 % para el otro, por lo que se obtiene un promedio de 65.23 % de exactitud.

Sin embargo, la exactitud es una métrica que toma en cuenta tanto los casos que coincidieron con el sistema al ser catalogados como co-hipónimos, como los que coincidieron con el corpus al ser catalogados como no-co-hipónimos, y este segundo conjunto no es proporcionado por nuestro sistema, de tal manera que para los propósitos de esta investigación es más relevante la precisión, es decir, de todos los pares que fueron extraídos como co-hipónimos, qué proporción es realmente percibida como co-hipónimos para los etiquetadores.

La precisión que se logra para el corpus basado en *BLESS* es de 67.55 % para uno de los etiquetadores, mientras que asciende hasta 69.29 % para el otro, con lo que se obtiene una precisión promedio de 68.42 %. Por el otro lado, en el corpus basado en *WordNet* la precisión aumentó a 63.71 % para uno de los etiquetadores, y hasta 69.88 % para el otro, con lo que se logra un promedio de precisión del 66.80 %. Este resultado muestra que, en realidad, la mayor proporción de desacuerdo entre etiquetadores se encuentra en las relaciones que fueron añadidas de *WordNet*; y en ambos casos, los etiquetadores coinciden en que casi el 70 % de los pares extraídos por nuestro método efectivamente se pueden considerar co-hipónimos. La tabla 7-1 muestra un resumen con los resultados que se obtuvieron con el etiquetado manual.

Para propósitos de comparación, este mismo proceso se repitió con un conjunto diferente

Tabla 7-1: Exactitud y precisión de los pares de co-hiponimia extraídos de forma automática. Se toma como *gold standard* las categorías ofrecidas por los etiquetadores humanos.

Corpus		Exactitud (%)	Precisión (%)
BLESS	Etiquetador-1	80.00	67.55
	Etiquetador-2	82.24	69.29
WordNet	Etiquetador-1	71.26	63.71
	Etiquetador-2	59.21	69.88

de co-hipónimos. Los patrones simétricos son, en nuestra opinión, la metodología más cercana a la que se está proponiendo, por tanto se usan como punto de comparación en un etiquetado equivalente.

Para este nuevo conjunto de co-hipónimos el acuerdo entre etiquetadores se mantuvo muy similar en ambos corpus. El porcentaje de acuerdo en el corpus basado en *BLESS* fue de 87.17 % con una κ de Cohen de 0.6846, mientras que el corpus basado en *WordNet* tuvo un acuerde de 64.63 % y una κ de 0.2790.

Una vez más, la evaluación de mayor importancia para los propósitos de esta investigación radica en la precisión, es decir, cuántos de los pares de palabras que se extrajeron mediante patrones de simetría son realmente percibidos como co-hipónimos. La precisión en el corpus basado en *BLESS* es de 58.82 % para uno de los etiquetadores, y de 62.70 % para el otro, por lo que se tiene un promedio de 60.76 %. Mientras que en el corpus basado en *WordNet*, la precisión fue de 57.94 % para uno de los etiquetadores y de solo 47.66 % para el otro, con lo que se obtiene un promedio del 52.80 %. La tabla 7-2 muestra un resumen de los datos antes mencionados. Como se puede observar, todos los resultados son sustancialmente inferiores a los que se logran con el sistema que se propone en este trabajo, con una diferencia de alrededor de 8 % de precisión.

Otro factor interesante que vale la pena considerar está en la intersección de las palabras que se consideraron como co-hipónimas. Es decir, aquellos pares de palabras que fueron percibidos como co-hipónimos para al menos uno de lo etiquetadores. Si estos casos se toman en cuenta

Tabla 7-2: Precisión para la percepción de co-hiponímia de pares extraídos mediante patrones simétricos

Corpus		Precisión (%)
BLESS	Etiquetador-1	58.82
	Etiquetador-2	62.70
WordNet	Etiquetador-1	57.94
	Etiquetador-2	47.66

como correctos en una evaluación general, combinando ambos etiquetados, la precisión del sistema que estamos proponiendo aumenta hasta 80.49 %. Si bien, este tipo de evaluación es más laxa, demuestra que los pares que se logran extraer son en su mayoría co-hipónimos, e incluso los que no, mantienen una relación muy similar a la co-hiponimia. En contraste, podemos comparar ese incremento (de más de 10 %) de precisión al calcular esta misma precisión laxa con los pares extraídos mediante patrones simétricos, en cuyo caso la precisión solo aumenta a un 66.69 % (menos de 6 % de diferencia), una mejora considerablemente menor, lo que se puede interpretar como que aquellos pares que no son co-hipónimos, también se encuentran más lejos de esta relación. Los detalles lingüisticos del umbral sobre el que algo comienza y deja de considerarse un co-hipónimo quedan fuera del alcance de esta investigación.

7.2. Antónimos

Como se mencionó con anterioridad, los pares de palabras muestran una tendencia a extraer co-hipónimos. No obstante, en la sección 5.3 se mostró cómo, a partir de un conjunto de co-hipónimos, se puede extraer un conjunto candidato de antónimos. Cabe recordar que, entre antónimos y co-hipónimos, la diferencia recae en la propiedad binaria de los antónimos, lo que les da su propiedad de ser opuestos. Además de los resultados de la binarización que se describieron en la sección 5.3, es necesario también analizar el aumento de la oposición que se logra con nuestro método para poder clasificar los resultados como antónimos completos

(oposición sistémica) o si se obtiene solo oposición no-sistémica. Por lo tanto, se llevó a cabo una evaluación similar a la que fueron sometidos los co-hipónimos. Los candidatos a antónimos fueron clasificados con el apoyo de dos equipos de etiquetadores humanos.

Dicho eso, la tarea que se le dio a los etiquetadores fue exactamente la misma que cumple el sistema: Clasificar la lista de co-hipónimos, la lista que se obtiene como salida del módulo de simetría (ver sección 5.2), como oposición sistémica, oposición no-sistémica, o ninguna. Para evitar el sesgo, un equipo de etiquetadores se dedicó a la clasificación de oposición sistémica mientras que el otro se dedicó a la clasificación de oposición no-sistémica. Esto nos provee de dos puntos de comparación para los pares extraídos de forma automática.

El resultado de este experimento nos permite evaluar el grado de oposición que se percibe por humanos con respecto a los pares de palabras que extrae nuestro sistema. El equipo de etiquetado de oposición sistémica, la cual es una clase muy restrictiva de antónimos, tuvo un porcentaje de acuerdo del 75.47 % con un coeficiente κ de Cohen de 0.4234, lo que equivale a un acuerdo moderado. Este equipo obtuvo como resultado que, en promedio, casi una tercera parte de los pares (30.61 %) extraídos son percibidos como antónimos puros (oposición sistémica).

Para el caso de la oposición no-sistémica, la cual es naturalmente una forma mucho menos restrictiva de antonimia, los etiquetadores tuvieron un porcentaje de acuerdo del 72.30 % con un coeficiente κ de 0.4232, casi el mismo acuerdo que el equipo de oposición sistémica. Este equipo obtuvo que, en promedio, el 39.73 % de los pares que se extrajeron podían ser considerados como antónimos de forma ocasional (oposición no-sistémica).

En la tabla 7-3 se muestra un resumen con los resultados de ambos casos.

Tabla 7-3: Porcentaje de precisión para la coincidencia de oposición sistémica y no sistémica con los pares extraídos de forma automática.

	Sistémica (%)	No-sistémica (%)
Etiquetador-1	28.91	37.19
Etiquetador-2	32.30	42.26
Promedio	**30.61**	**39.73**

Es interesante observar que la diferencia de percepción entre la oposición sistémica y la oposición no-sistémica no es demasiado grande, esto se puede tomar como un indicador de que el filtro por binarización está funcionando proporcionalmente más para la oposición sistémica que para la no-sistémica. Lo anterior sigue cierta lógica, pues cuando se tiene oposición sistémica, la cantidad de palabras con las que se puede emparejar una palabra opuesta es mayor (ya que depende de su contexto). Los resultados de la evaluación nos demuestran que la binarización efectivamente ayuda para la extracción de oposición. Sin embargo, es apenas el primer paso, pues aún queda mucho espacio para mejorar, y la mayoría de los pares que se extraen, aunque sí son percibidos como co-hipónimos (ver evaluación de la sección 7.1), aún faltan mejores filtros para lograr mejores antónimos.

7.3. Enriquecimiento de *embeddings*

Para continuar con el análisis de los resultados y sobre todo para poder evaluar la contribución que puede tener la oposición que se obtiene en la práctica, se replicaron los experimentos de Mrkšić *et al.* [2016] para enriquecer representaciones distribucionales mediante la adición de información de oposición. El enriquecimiento se llevó a cabo para ambos casos, tanto para los co-hipónimos como para los antónimos. Al igual que en los experimentos de Mrkšić *et al.* [2016], se usó *counter-fitting* para agregar restricciones lingüísticas a GloVe *embeddings* pre-entrenados.

Faruqui *et al.* [2014] diseñaron *retrofitting*, un método para refinar los vectores de un espacio vectorial (*word embeddings*) de tal manera que se pudiera unir la información de relaciones de lexicones (como *WordNet*); y todo sin la necesidad de dar a los vectores una estructura especial o un espacio vectorial ajeno al original. El *retrofitting* es una técnica basada en grafos que se aplica como post-procesamiento para obtener vectores de mayor calidad semántica. El principal objetivo es que los nuevos vectores sean similares tanto a los vectores originales como a los vectores de las palabras relacionadas según los recursos externos; para lograrlo, se calcula la distancia euclidiana entre los pares de los vectores de tal manera que la función objetivo queda de la forma:

$$\hat{J} = \sum_i \left(\alpha_i ||q_i - \hat{q}_i||^2 + \sum_{(i,j) \in E} \beta_{ij} ||q_i - q_j||^2 \right)$$

donde q_i pertenece al nuevo espacio de vectores a encontrar: (Q); y $\hat{q}_i$ pertenece al espacio vectorial original. $(\hat{Q}$; Los pares E (i, j) pertenecen a las palabras relacionadas en los recursos exteriores; y tanto α como β representan hiper-parámetros de control para la ponderación relativa de los términos. Gracias a este método, los autores logran mejoras en el desempeño de tareas como WS-353 y la detección de sinónimos para TOEFL.

Usando como base el *retrofitting*, Mrkšić *et al.* [2016] propusieron un nuevo sistema para integrar información a sistemas de *embeddings*, pero esta vez de información contrastante. El método, llamado *counter-fitting*, consiste en transformar el espacio vectorial V a un espacio V' modificado después de agregar información tanto de sinónimos como de antónimos; el método consiste en tres partes:

Repulsión de antónimos (AR):

Este es el término responsable de incrementar la distancia entre antónimos dentro del espacio vectorial, la transformación queda de la forma:

$$AR(V') = \sum_{(u,w) \in A} \tau \left(\delta - d(v'_u, v'_w) \right)$$

donde:

$d(v_i, v_j) = 1 - \cos(v_i, v_j)$ es una distancia que se deriva del coseno para la similitud

$\tau(x) = \text{máx}(0, x)$ se usa para imponer un margen en el costo

δ es la distancia mínima "ideal" entre palabras antónimas

A es el conjunto de palabras antónimas

Atracción de sinónimos (SA):

Este es el término responsable de reducir la distancia entre pares de sinónimos para el

nuevo espacio vectorial, la transformación queda de la forma:

$$SA(V') = \sum_{(u,w)\in S} \tau\left(d(v'_u, v'_w) - \gamma\right)$$

donde:

γ es la distancia máxima "ideal" entre pares sinónimos de palabras

S es el conjunto de palabras sinónimas

Preservación del espacio vectorial (VSP):

Este término es el encargado de garantizar que las relaciones que fueron capturadas en el entrenamiento original se mantengan tanto como sea posible a pesar de los cambios de los términos anteriores, de esta forma se garantiza que el espacio vectorial transformado conserve su utilidad y proporción general.

$$VSP(V, V') = \sum_{i=1}^{N} \sum_{j\in N_{(i)}} \tau\left(d(v'_i, v'_j) - d(v_i, v_j)\right)$$

donde:

$N_{(i)}$ es el conjunto de palabras que queda dentro de un radio ρ alrededor de una palabra i en el espacio original V

La función objetivo final de *counter-fitting* se da mediante una suma ponderada de los tres términos mediante los hiper-parámetros k_1, k_2 y k_3:

$$C(V, V') = k_1 AR(V') + k_2 SA(V') + k_3 VSP(V, V')$$

Gracias a esta técnica, es posible evaluar la utilidad que representa la información de oposición que se extrae en los resultados de esta investigación mediante la medición del cambio en el desempeño de un sistema sobre tareas que requieran distinguir entre palabras similares y palabras asociadas. No obstante, es importante aclarar que para los propósitos de este trabajo,

solamente es relevante la contribución que se logra gracias al conjunto de palabras antónimas, por lo que las restricciones lingüísticas de sinonimia fueron omitidas en la evaluación.

Para contar con un punto de referencia y comparación se utilizaron, al igual que en el trabajo de Mrkšić *et al.* [2016], dos recursos externos como fuente de restricciones lingüísticas, en adición a los resultados que se extrajeron automáticamente en este trabajo.

***WordNet*:** Un lexicón anotado manualmente con relaciones como la sinonimia, antonimia, hiperonimia, etc. [Miller, 1995]. En el experimento, todos los pares de palabras que están etiquetados como antónimos se usaron para el enriquecimiento de *embeddings*

PPDB 2.0: Un corpus para paráfrasis que cuenta con alineaciones de palabras [Pavlick *et al.*, 2015]. En particular, la relación de *Equivalence* se toma como sinónimo y la relación de *Exclusión* como antónimo en los experimentos de Mrkšić *et al.* [2016].

Mrkšić *et al.* [2016] utilizan en sus experimentos los dos corpus anteriores para obtener relaciones de antónimos y exclusión. Al enriquecer GloVe con esta información logran mejoras de desempeño para la tarea SimLex-999. SimLex-999 es un *gold standard* para la evaluación de modelos que obtienen representaciones del significado de las palabras. La principal diferencia entre SimLex-999 y otros corpus similares, es que en este caso se hace una clara distinción entre palabras asociadas y palabras similares, y solamente se le asignan valores altos a palabras con un alto grado de similitud semántica, y valores bajos a aquellas que solo están asociadas. Al atraer entre sí a los sinónimos, y al repeler entre sí a los antónimos, Mrkšić *et al.* [2016] buscan incrementar la similitud de palabras efectivamente similares mientras desvanecen la similitud presente en palabras que solo están asociadas, este enfoque funciona bien en su artículo.

Considerando lo anterior, en esta investigación se retoma esta misma idea y se enriquecen los *embeddings* de *GloVe* mediante *counter-fitting* para repeler palabras opuestas. Como se mencionó anteriormente, se omitió la inclusión de sinónimos para apreciar de mejor manera los cambios que se producen tan solo por oposición. Además de las dos fuentes externas que toman Mrkšić *et al.* [2016] y nuestras propias dos fuentes de pares de palabras (co-hipónimos y antónimos), se incluyen también en los experimentos los pares de palabras extraídos mediante

patrones simétricos [Schwartz *et al.*, 2015]. Como se ha mencionado antes, consideramos que esta metodología de extracción de relaciones léxicas es la más cercana a esta investigación.

En la tabla 7-4 se muestran los resultados del experimento, donde se comparan las diferentes fuentes de restricciones lingüísticas, así como varias de sus combinaciones. Dichas restricciones se muestran enumeradas a continuación:

-: El primer resultado es la *baseline* de comparación, se utilizaron vectores pre-entrenados de *Glove* sin ninguna modificación. Todos las restricciones lingüísticas fueron aplicadas sobre estos mismos *embeddings*.

PPDB: Los antónimos extraídos del PPDB 2.0 al tomar la relación *Exclusion* como antónimo.

***WordNet*:** Pares de palabras identificadas como antónimos en *WordNet*

***SimAnts*:** Pares de palabras consideradas como antónimos para Schwartz *et al.* [2015]. Estos antónimos se extraen mediante los patrones simétricos "*from* X *to* Y" y "*either* X *or* Y"

***SimFull*:** Todos los pares de palabras extraídas con los patrones simétricos reportados por Schwartz *et al.* [2015].

***Yux-Cos*:** Todos los pares de co-hipónimos extraídos en este trabajo siguiendo la metodología hasta la sección 5.2.

***Yux-Ants*:** Pares candidatos a antónimos extraídos a partir de los co-hipónimos, el resultado final de toda la metodología (sección 5.3).

Se puede apreciar en los resultados cómo hay un incremento en el desempeño del sistema cuando se incrementa la separación de la representación de palabras "antónimas". Incluso si dichos pares de palabras no son antónimos en el sentido más estricto de la palabra. Para facilitar la visualización, en la figura 7-2 se muestran de manera gráfica las ganancias que logran cada una de las restricciones lingüísticas de forma individual.

El corpus *PPDB* cuenta con tan sólo 478 pares de palabras, entre las cuales aparecen muchos co-hipónimos (como se puede observar en la tabla 5-1). A pesar de ello, logra un

Tabla 7-4: Resultados de la evaluación *SimLex999* al aplicar *counter-fitting* sobre *embeddings Glove* con una variedad de restricciones lingüísticas.

Restricción	**Num. de pares**	ρ
-	0	0.418
PPDB	478	0.445
WordNet	12,470	0.528
SimAnts	8,396	0.451
SimFull	809,372	0.205
Yux-Ants	5,346	**0.482**
Yux-Cos	13,764	**0.44**
SimAnts+*PPDB*	8,764	0.473
Yux-Ants+*PPDB*	5,728	**0.487**
WordNet+*PPDB*	12,802	0.53
Yux-Ants+*SimAnts*	13,368	**0.509**
Yux-Ants+*SimAnts*+*PPDB*	13,676	**0.509**
WordNet+*SimAnts*	20,510	0.538
WordNet+**Yux-Ants**	17,554	**0.57**
WordNet+**Yux-Ants**+*SimAnts*	25,296	**0.573**

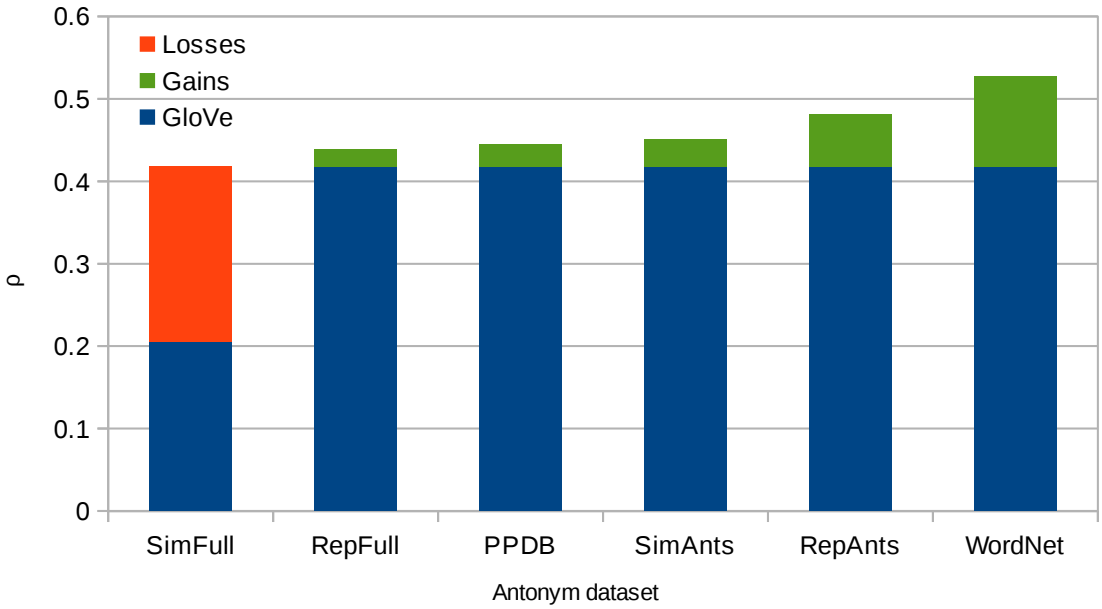

Figura 7-2: Correlación de Pearson para cada una de las fuentes de restricciones lingüísticas sobre la tarea *SimLex-999*. Se muestran resaltadas las ganancias y pérdidas que implica cada una de estas fuentes externas.

aumento de 0.418 a 0.445 de correlación de Pearson. No obstante, 478 representa una cobertura bastante pobre y al combinarlos con otras fuentes de información su aportación se vuelve poco representativa y no aporta ninguna ganancia a la combinación de los métodos **Yux-Ants** con *SimAnts*, como se puede observar en la tabla 7-4.

Los sistemas *Yux** y **Sim*** (los resultados de esta investigación y los que usan patrones simétricos para la extracción de pares respectivamente) aumentan la cantidad de pares de palabras a diferenciar mediante patrones léxicos. En el caso de los sistemas simétricos (*Sim**) se extraen los pares mediante los patrones simétricos reportados por Schwartz *et al.* [2015]. El sistema *SimFull* considera todos los pares que se extraen mediante dichos patrones simétricos. Sin embargo, como se puede observar, debido a la gran cantidad de pares que se extraen, se puede asumir que la precisión de la información que aportan como antonimia o co-hiponimia es muy baja, al igual que la evaluación. Los pares que se extraen con este método son muy ruidosos y predomina la información circunstancial, son más aptos para ser tomados como ventanas de contexto (tal cual lo hacen los autores). Este puede ser el motivo por el cual, a

pesar de que muchos de estos patrones se han encontrado presentes con muchas palabras de oposición [Mettinger *et al.*, 1994; Justeson y Katz, 1991; Fellbaum, 1995; Jones, 2003], nunca han sido usados como un método de extracción. Schwartz *et al.* [2015] mencionan además en su artículo que usaron los patrones: *"from X to Y"* y *"either X or Y"* para detectar antónimos con mayor precisión. Para el sistema *SimAnts* se tomaron solo los pares de dichos patrones para ofrecer un punto de comparación mucho más depurado. En nuestros experimentos este sistema supera los resultados del *PPDB*.

Por otro lado, los sistemas **Yux*** se extraen mediante nuestra metodología. El sistema *Yux-Cos* considera todos los pares de palabras co-hipónimos que se extraen siguiendo el método hasta la sección 5.1. Como se puede apreciar en la tabla, estos pares representan una mejora sobre el sistema original, lo que indica que en su mayoría de hecho aportan información útil. Sin embargo, a pesar de extraer más pares que *Yux-Ants*, tiene un peor desempeño, lo cual es un indicador de que en este caso también se extraen pares ruidosos con información circunstancial para la tarea en cuestión.

En la figura 7-3 se muestran las ganancias que se obtienen al aplicar los pares de co-hipónimos extraídos automáticamente (*Yux-Cos*) en combinación con los otros sistemas externos que logran mejorar el rendimiento de la *baseline*. Como se puede observar, la diferencia que se obtiene, si bien es pequeña, es positiva en todos los casos. Esto quiere decir que incluso, al tener intersección en los (relativamente) pocos pares que mejoran el desempeño, los otros no logran perjudicarlo; en general, tampoco se están capturando como opuestos muchos elementos que se deban considerar como semejantes.

Gracias a la adición de binarización en la sección 5.3, la metodología llegó a la altura de aplicar un filtro completamente no supervisado para la obtención de pares de palabras con gran nivel de oposición. Estos pares de palabras se reportan como ***Yux-Ants*** y muestran mucho mayores beneficios en el desempeño que los pares ***Yux-Cos*** a pesar de extraer menos pares. Eso demuestra la mayor precisión y confianza que poseen para beneficiar aplicaciones de esta naturaleza.

Tan solo por debajo de los antónimos de *WordNet*, los pares extraídos por ***Yux-Ants*** son los

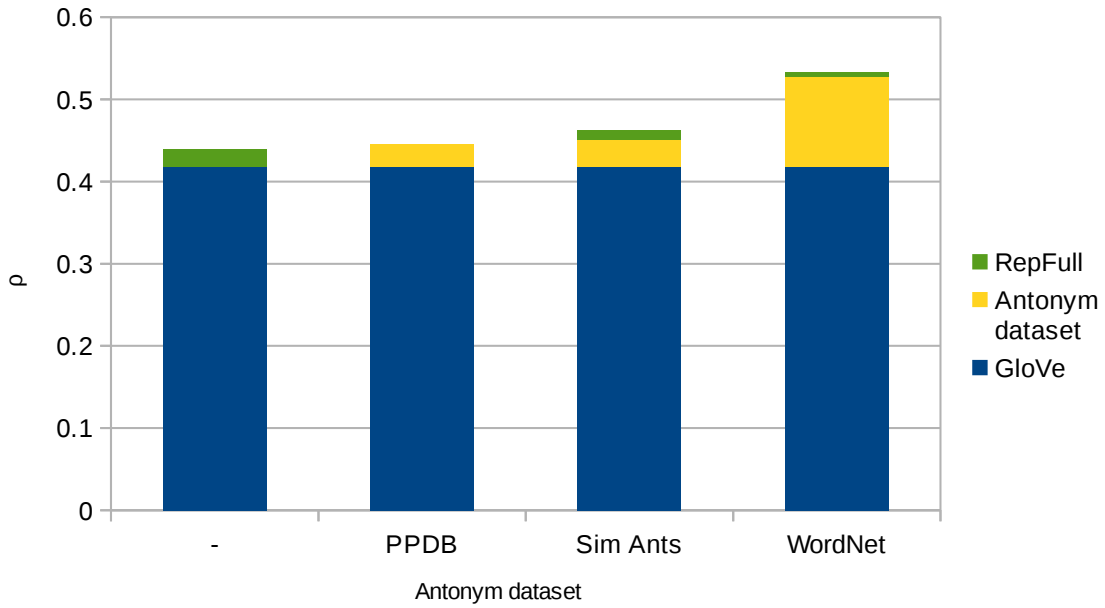

Figura 7-3: Ganancias en la correlación de Pearson sobre la tarea *SimLex-999* al aplicar los pares **Yux-Cos** de forma adicional a las otras fuentes de restricciones lingüísticas.

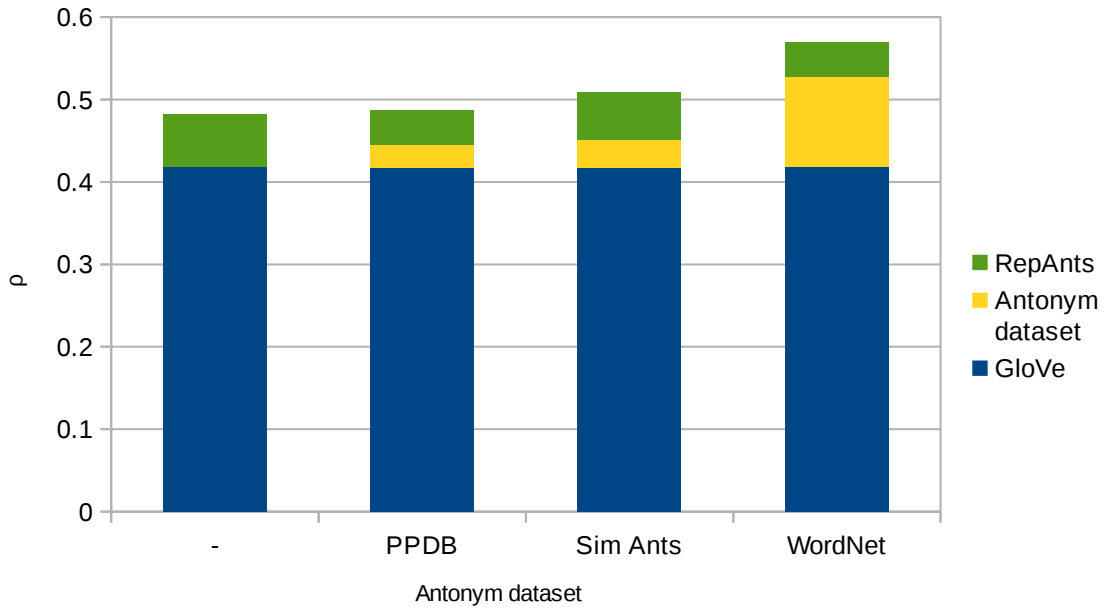

Figura 7-4: Ganancias en la correlación de Pearson sobre la tarea *SimLex-999* al aplicar los pares **Yux-Ants** de forma adicional a las otras fuentes de restricciones lingüísticas.

que presentan mayor margen de ganancia, tanto en solitario como al ser combinados con otros sistemas. Esto último también es prueba de su capacidad para aportar información relevante que los otros sistemas pueden pasar por alto. En la figura 7-4 se muestra de forma gráfica el margen de ganancias que se logra al añadir estos últimos pares que toman en cuenta la binarización en combinación con los otros sistemas externos que lograr mejorar el rendimiento de la *baseline*.

Cuando comparamos las figuras 7-3 y 7-4, podemos observar que el último filtro de binarización que se aplica a los co-hipónimos (ver sección 5.3) presenta una contribución importante cuando se evalúa la efectividad que representa esa oposición en tareas de esta naturaleza, incluso cuando en el etiquetado manual (ver sección 7.2) se percibe oposición fuerte en tan solo 30 % de los casos. Esta última evaluación muestra una mejora importante a pesar de que se tienen relativamente pocos pares de palabras extraídos. Si se toma la ganancia por número de pares, la calidad de oposición de nuestros "antónimos" es comparable con los antónimos de *WordNet*.

De forma complementaria a los experimentos anteriores, se efectuó también un experimento similar pero usando como restricciones lingüísticas los resultados que se obtuvieron a partir del corpus de español y, por supuesto, con la versión de GloVe pre-entrenada de español [Pennington *et al.*, 2014; Cardellino, 2019].

Para la evaluación usamos dos corpus en español que fueron construidos en la investigación de Hassan y Mihalcea [2009]. El primero de dichos corpus fue basado en el corpus WS353 [Finkelstein *et al.*, 2001], que consiste en un corpus de 353 pares de palabras etiquetado por 13 expertos humanos en una escala del 0 al 10. El segundo se basó en el MC30 [Miller y Charles, 1991], un corpus que consiste en 30 pares de palabras etiquetadas, cada una por 38 humanos con una escala del 0 al 4. Los resultados de los experimentos se pueden observar en la tabla 7-5.

Tabla 7-5: Resultados del enriquecimiento con oposición sobre *embeddings GloVe*.

Corpus	ρ **original**	ρ **enriquecida**
WS353	0.333	0.391
MC30	0.467	0.61

En este caso nos encontramos ante una tarea en la que no se han desarrollado tantos sistemas

ni puntos de comparación en el español como los hay para el inglés. No obstante, incluso en estos casos podemos ver una mejora en los resultados al aplicar un enriquecimiento de oposición sobre los *embeddings* pre-entrenados de una forma similar a como ocurre con el inglés, e incluso se logran mayores cambios en el rendimiento particular. Por supuesto, se trata de corpus de evaluación diferentes y es muy pronto para sacar conclusiones generales sobre la tarea. Pero el objetivo del experimento es conocer la aportación que puede tener la adición de información de oposición sobre *embeddings* pre-entrenados. Así, se pueden observar mejoras, tanto en español como en inglés, que reflejan una cierta independencia del idioma.

7.4. Representación de negación

En la sección 2.5 introdujimos el concepto de la negación y la forma en la que se relaciona con la oposición. En esta sección retomaremos esta idea, pues buscaremos representar la negación con apoyo de los resultados de nuestro método para evaluar la aportación que puede tener sobre una aplicación práctica, en este caso, sobre un sistema de minería de opinión. Parte de las ventajas que se presentaron a lo largo de la metodología fue la relativa independencia del idioma del método, lo cual se observó a lo largo del capítulo 6. Esta sección de representación de negación retoma los resultados del español y analizamos la tarea en este idioma.

Para este tipo de experimentos es necesario partir desde una estructura para expresar la negación dentro de un texto. En este trabajo, utilizamos el esquema que se presenta en Bel-Enguix *et al.* [2021], el cual, a su vez, toma como base el corpus SFU Review$_{\text{SP}}$-NEG [Jiménez-Zafra *et al.*, 2018] como punto de comparación para la evaluación de la tarea. El corpus consiste de 400 revisiones y comentarios extraídos de *Ciao.es*, que ascienden a un total de 9,455 oraciones, de las cuales 3,022 contienen al menos una negación. El esquema de etiquetado es importante, ya que se cuenta con una identificación no solo del marcador de negación, sino del evento y, en general, del *scope* sobre el que tiene efecto la negación.

En la figura 7-5 se muestra la estructura general de los corpus con etiquetado de negación. Se puede observar que para cada texto se tiene la posibilidad de marcar una sección en la que hay

una negación presente. Dicha sección de texto está compuesta por una estructura de negación, es decir, una expresión o marcador negativo (por ejemplo: no, ni, ningún, etc.), un evento el cual está siendo negado y el *scope* completo en el que la negación tiene efecto [Bel-Enguix *et al.*, 2021].

```
<tweet>
    <content>
        <neg_structure>
            <scope>
                <negexp class='simple/related/no_neg'>
                </negexp>
                <event>
                </event>
            </scope>
        </neg_structure>
    </content>
</tweet>
```

Figura 7-5: Estructura general del etiquetado de negación.

Gracias a esta estructura de la información de negación, es posible plantear un experimento de evaluación de la oposición. En trabajos como Shah y Rekh [2014] se resalta la importancia de tomar en cuenta la negación para tareas como la de minería de opinión; esto es debido al cambio diametral de significado que una negación implica a lo que se está diciendo; es decir, una oración que normalmente se podría considerar positiva o negativa cambia de polaridad al ser sujeto de una negación, por ejemplo:

El celular **no** tiene una buena calidad de llamada, pero su batería **no** dura poco.

Tenemos nuevamente una opinión similar a la que se planteó en la sección 2.5, pero se han añadido los "**no**", por lo que ahora ambas partes están negadas. Antes teníamos una opinión positiva para la calidad de la llamada, mientras que se tenía una opinión negativa para la batería, ahora la polaridad ha quedado completamente invertida. Esto se podría entender de una manera similar a que si estuviera escrito como:

El celular tiene una **mala** calidad de llamada, pero su batería dura **mucho**.

Es decir, el "**no**" representa una transformación de las palabras originales hacia una versión opuesta de las mismas. Es justo aquí donde se puede hacer un enriquecimiento gracias a los

resultados de esta investigación, ya que podemos buscar obtener dicha transformación usando como base la oposición que se logró extraer.

Dado que la transformación se hace a nivel de las palabras, se plantea nuevamente el uso de *word embeddings* para obtener una representación vectorial a la cual se le aplicará la transformación negativa. En otras palabras, si se tiene una representación vectorial (v) de dimensión d, se buscará obtener una matriz N (matriz que representa una transformación de negación) de dimensión $d \times d$, para que al ser aplicado al vector v se obtenga una versión "negada" de la palabra (v').

Para la obtención de la matriz de transformación N, se entrenó una red neuronal lineal sin función de activación, en el que se usaron todos los vectores de los pares de palabras que resultaron de la sección 5.3 aplicada al español, como entrada, y sus respectivos pares, como salida. Dicha matriz de transformación se usó para la aplicación de minería de opinión.

Para evaluar la aportación, se desarrolló un sistema sistema de clasificación de sentimientos para las entradas de textos de opiniones. El sistema recibe los textos etiquetados con negación, como se ha referido anteriormente, y obtiene una representación a partir de los *embeddings* de sus palabras. Particularmente, se usa una combinación lineal ponderada de los vectores en la que los pesos se ajustan durante el entrenamiento. De esta manera, será posible comparar un sistema en el que los vectores son alimentados sin ninguna modificación, y un sistema en el que aquellas palabras etiquetadas como parte del *scope* de una negación (excluyendo la expresión de negación, la cual se elimina de la entrada) son modificadas previamente agregando la transformación de oposición previamente entrenada (N). Adicionalmente, el mismo experimento se repite usando como *embeddings* una versión modificada de GloVe, usando *counter-fitting* [Mrkšić *et al.*, 2016], de la misma manera a como se hizo para la evaluación de la sección 7.3.

Con estos procesos obtuvimos los resultados que se muestran en la tabla 7-6. Como elementos de comparación, se utiliza un sistema en el que como transformación del *embedding* de entrada se usa el vector negativo, y se usan también los resultados que se han obtenido de una investigación paralela de minería de opinión del Grupo de Ingeniería Lingüística[1], los cuáles se identifican

[1]Se agradece de forma particular la colaboración del integrante Brian Aguilar Vizuet.

como se indica a continuación:

BOW: Un sistema que usa una representación de bolsa de palabras para los documentos y un clasificador que utiliza SVM.

LEX: Un sistema que hace uso del lexicón especializado para sentimientos. [Chen y Skiena, 2014]. Las representación de las oraciones se basa en los sentimientos marcados por estas palabras clave.

SVM: Este sistema es similar al anterior, pues además del uso de lexicón especializado, también hace uso de una representación de bolsa de palabras y el sistema se entrena con un SVM lineal.

TF-IDF: Un sistema que combina el lexicón especializado ponderado mediante TF-IDF para la representación de los documentos.

GloVe: Este es el sistema base que se usó con el método que se expuso en esta sección, es decir, una red neuronal que usa como representación de las oraciones una combinación lineal ponderada de las palabras de entrada, cada palabra se obtiene de *GloVe* pre-entrenado [Pennington *et al.*, 2014].

I-NEG: Este se trata de un sistema comparativo en el que se aplica una transformación al vector de inversión de signo, en lugar de la transformación lineal que se obtiene con la oposición.

M-NEG: Este es el sistema que se propone en el que aquellas palabras que se ubican dentro del *scope* de una negación son multiplicadas por una matriz de transformación lineal que se obtiene a partir de los pares de oposición de esta investigación.

COUNTER: El último sistema que reportamos es una modificación del M-NEG, en el que los *embeddings* que se usan son la versión modificada mediante *counter-fitting* [Mrkšić *et al.*, 2016], por lo que tanto la representación original de las palabras como la matriz de

transformación de negación quedan modificadas de la misma manera que en la metodología de la sección 7.3.

Tabla 7-6: Resultados de diferentes sistemas de clasificación de sentimiento.

Sistema	*Precision*	*Recall*	*F-score*
BOW	0.78	0.19	0.31
LEX	0.78	0.58	0.66
SVM	0.63	0.51	0.52
TF-IDF	0.71	0.7	0.71
GloVe	0.68	0.61	0.64
I-NEG	0.7	0.58	0.63
M-NEG	0.76	0.7	0.73
COUNTER-full	0.71	0.63	0.67

Estos resultados presentan datos interesantes para ser analizados. En primer lugar, podemos ver la importancia del uso de lexicones especializados en sentimiento, de manera que el sistema **LEX** logra resultados semejantes a sistemas mucho más complejos, usando únicamente esa fuente de información.

Otro dato importante es el obtenido por nuestro sistema **M-NEG**, el cual resaltamos, ya que logra los mejores resultados; sin embargo, el resultado es equiparable con el segundo lugar, **TF-IDF**, un sistema que combina *TF-IDF* con el lexicón y la información de negación. No obstante, nuestro sistema logra dichos resultados sin recurrir al conocimiento externo del lexicón y, más importante, sí representa una mejora con respecto a los sistemas que no hacen la transformación lineal que se plantea en esta sección.

Por último, un resultado que también resulta muy interesante es el del sistema **COUNTER**, ya que obtiene una mejora mínima, a pesar de usar la misma base de entrenamiento de transformación lineal. Este resultado sugiere que, si bien se logran representaciones generales más apegadas a la percepción como se vio en la evaluación de la sección 7.3, el *counter-fitting* [Mrkšić *et al.*, 2016] en el proceso ocasiona una pérdida de la relación general espacial que tienen

los términos con sus opuestos. En este punto, podemos teorizar la razón de que esto ocurra, debido a la naturaleza del proceso de *counter-fitting*: es posible que el alejamiento general de los elementos opuestos dentro del espacio vectorial ocasionen deformaciones heterogéneas que, al intentar ser capturadas por una matriz, se neutralicen entre ellas. Sin embargo, la exploración estricta de este comportamiento queda fuera del alcance de esta investigación.

Capítulo 8

Conclusiones

Una meta ideal del procesamiento automático del lenguaje es lograr, para una computadora, la abstracción artificial de lo que lee, es decir, que una máquina tenga una representación que pueda manejar el significado de lo que se presenta en un documento. En el presente trabajo se buscó contribuir a esta meta a través de aprovechar la extensa disponibilidad de texto con la intención de encontrar relaciones propias de una base léxica para enriquecer tanto dichas bases léxicas como un sistema de *embeddings*.

Se tiene un gran potencial de enriquecimiento de sistemas basados en representaciones distribucionales debido a la forma en la que calcula la asociación entre las palabras que tiende a asociarse más con la sinonimia que con cualquier otra relación léxica (y principalmente lejana a relaciones de oposicón). Ésta área de oportunidad responde a una propiedad intrínseca de la semántica distribucional y es la razón por la que la oposición es la relación que a estos sistemas más les cuesta distinguir.

Para dicho propósito, se llevó a cabo el diseño de una metodología que busca encontrar palabras relacionadas por oposición de manera que se puedan tomar como fuente de información, interna y propia de los textos, adicional a la ya usada co-ocurrencia. Ello, sin la necesidad de recurrir a recursos lingüísticos externos y manteniendo, en la medida de lo posible, la independencia del idioma y de la necesidad de recursos costosos para su funcionamiento.

Para atender el enriquecimiento automático de recursos léxicos, se llevó a cabo la identi-

ficación de los distintos tipos de relaciones paradigmáticas entre las palabras, principalmente antónimos. Experimentos preliminares sobre los textos ofrecieron indicios que sugerían un fenómeno interesante: el uso de palabras relacionadas, pero contrapuestas dentro de ventanas de palabras funcionales particulares. Muchas de dichas ventanas mantienen, para ambas palabras, la misma estructura funcional, lo que da lugar a un patrón de repetición o yuxtaposición.

Lo anterior nos llevó a pensar que era factible dividir a las palabras vacías en dos tipos de funcionales según el lugar en el que suelen aparecer en los patrones yuxtapuestos, ya sea como acompañantes de las palabras de contenido (artículos principalmente) o como palabras que conectaran las dos partes de la yuxtaposición (preposiciones y conjunciones entre otras). Esta división y forma de separar las palabras vacías permitieron llevar a cabo el primer módulo de la metodología: un proceso de extracción de patrones yuxtapuestos sin la necesidad de contar ni siquiera con la lista de palabras funcionales. Las propiedades de dicho módulo contribuyeron en el propósito de mantener el método relativamente independiente del idioma y que funcionara también para idiomas de bajos recursos y contextos especializados, por ejemplo en redes sociales, donde se cambian muchas palabras y las funcionales típicas podrían no ser suficientes.

A partir de la extracción de patrones yuxtapuestos, la metodología continuó con el refinamiento de pares de palabras para detectar y obtener aquellos que presentaran mayor grado de oposición. Para ello, en el segundo módulo se usó la simetría para filtrar aquellas relaciones asimétricas (y que por tanto no eran de oposición) dentro de las que se extrajeron en el primer módulo. Los patrones yuxtapuestos dieron lugar a diferentes palabras vacías conectoras, cada una con un grado distinto de simetricidad; este dato también se usó para no tomar en cuenta las conexiones con una gran caída de rango frecuencia-simetricidad, caída característica de conexiones genitivas y, por tanto, demasiado ruidosas para ser aprovechadas.

Por último, se tuvo un módulo de binarización. La salida del módulo anterior mostró mucha similitud en relaciones de palabras con la co-hiponimia, pero para alcanzar un grado de oposición similar al de la antonimia fue necesario un paso de refinamiento final. Los antónimos son elementos extremos de co-hipónimos binarios. Siguiendo esa lógica, se diseñó el último módulo del método para extraer todos aquellos pares relacionados con solo una palabra. Este último

filtro dio como resultado los pares con mayor grado de oposición y candidatos a ser considerados como antónimos.

La aplicación de la metodología mostró su capacidad de extraer estos patrones yuxtapuestos de manera automática, de tal forma que se extrajeron pares de palabras con oposición (antónimos y co-hipónimos) a partir de un texto plano.

Ya que uno de los principales objetivos de este trabajo fue mantener, en lo posible, el sistema lo más independiente del idioma, los experimentos que se realizaron para el el español se repitieron para el inglés. Al repetir el experimento de los patrones de yuxtaposición, sin cambios, se obtuvieron también dos listas de palabras vacías que mantenían las mismas funciones que para el español. En los resultados obtenidos, se pudo notar una gran similitud con los resultados que se obtuvieron tanto en español como para el inglés sin la necesidad de cambiar en nada el algoritmo. Estos resultados son muy importantes, ya que demuestran cierto grado de independencia del idioma para el método.

En general, los resultados hasta este punto obtenidos son buenos indicadores de que la metodología funcionará en idiomas que se puedan tokenizar y usen pre o post-posiciones.

En relación con el cumplimiento de la hipótesis, se puede añadir que también se cumplió con el objetivo de presentar una metodología en que se mantuviesen al mínimo los recursos lingüísticos necesarios para la extracción de relaciones léxicas entre las palabras de un texto y de ahí generalizar los procedimientos y enriquecer las representaciones de palabras de una forma totalmente no supervisada. De esta manera, una contribución del presente trabajo es el de contar con herramientas en las que se pueda prescindir de recursos léxicos, sobre todo para situaciones en las que se trabaje con lenguas de bajos recursos o en contextos especializados. Lo anterior, aunado a que puede haber polémica en cuanto a qué se considera una palabra funcional, ha llevado al diseño del método aquí propuesto, el cual no requiere de conocer las palabras funcionales previamente, ya que la misma estructura es capaz de detectar e incluso clasificar palabras vacías en un texto sin necesidad de tener información a priori.

La evaluación de mayor importancia para los propósitos de esta investigación radica en la precisión, es decir, cuántos de los pares de palabras que se extrajeron mediante el método

propuesto son realmente percibidos como co-hipónimos. Se hizo una evaluación con el apoyo de dos corpus usados como contraejemplos: BLESS y WordNet. Así, la precisión en el corpus basado en BLESS fue de 60.76 %. Mientras que en el corpus basado en WordNet, la precisión fue del 52.80 %. Se hizo una comparación con pares obtenidos mediante un sistema basado en patrones simétricos, el método más cercano al que se propone, todos los resultados de esta *baseline* fueron sustancialmente inferiores a los que se logran con el sistema que se propone en este trabajo, con una diferencia de alrededor de 8 % de precisión. Otro factor interesante que vale la pena considerar está en la intersección de las palabras que se consideraron como co-hipónimas. Es decir, aquellos pares de palabras que fueron percibidos como co-hipónimos para al menos uno de lo etiquetadores. Si estos casos se toman en cuenta como correctos en una evaluación general, combinando ambos etiquetados, la precisión del sistema que estamos proponiendo aumenta hasta 80.49 por ciento.

Por otro lado, también se llevó a cabo una evaluación de enriquecimiento de representación vectorial de palabras. Para llevar a cabo esta etapa, se agregó la información de oposición de diferentes fuentes a un sistema pre-entrenado GloVe mediante el proceso de *counter-fitting*. El enriquecimiento se hizo tanto para los candidatos a co-hipónimos como con los candidatos a antónimos que se extrajeron mediante la metodología propuesta. Los pares obtenidos se compararon con pares extraídos mediante patrones simétricos, y también con pares extraídos de los recursos léxicos PPDB y WordNet.

Los co-hipónimos extraídos por patrones simétricos representaron pérdidas al ser combinados con el sistema de *embeddings* base ($\rho = 0.42$), pero todos los otros sistemas obtuvieron ganancias. Nuestros pares de co-hipónimos lograron una ganancia equivalente a los pares del PPDB (aumento de ρ en 0.03) aunque por debajo de los patrones simétricos selectos (patrones simétricos elegidos a mano para mejorar su extracción de antónimos con aumento de ρ de 0.04).

Por su parte, nuestros pares de antónimos lograron superar a todos los sistemas anteriores ($rho + 0.07$) y quedaron solo por debajo de los pares de WordNet ($rho + 0.11$). No obstante, es relevante mencionar que el conteo de pares de WordNet también fue mayor. Si se considera una ganancia por cantidad de pares, nuestros antónimos tienen la misma calidad de oposición

(para esta prueba) que los que se encuentran en WorNet. Más aún, quizá lo más importante es la combinación de los dos sistemas, es decir, el método que proponemos es capaz de aportar información que no está presente en WordNet y esto se demuestra ya que, al ser combinado con este, la ganancia subió nuevamente ($rho + 0.16$).

Gracias a estos experimentos se puede aseverar que la aplicación de la metodología mostró su capacidad de extraer pares de palabras con oposición (antónimos y co-hipónimos) a partir de un texto plano, tal que dichos pares son, en su mayoría, coincidentes con la percepción humana y además representan una fuente adicional capaz de enriquecer representaciones vectoriales en un ámbito que resulta complicado de adquirir para sistemas distribucionales.

Bibliografía

Abdalgader, K. Text-Fragment Similarity Measurement using Word Sense Identification. *International Journal of Applied Engineering Research* **11**(24):11755–11762 (2016)

Acosta, O., Aguilar, C., y Sierra, G. A method for extracting hyponymyhypernymy relations from specialized corpora using genus terms. En *Proceedings of the Workshop in Natural Language Processing and Web-based Technologies*, págs. 1–10 (2010)

Acosta, O., Aguilar, C.A., y Sierra, G. Using Relational Adjectives for Extracting Hyponyms from Medical Texts. En *AIC@ AI* IA*, págs. 33–44. Citeseer (2013)

Acosta, O., Sierra, G., y Aguilar, C. Extracting definitional contexts in Spanish through the identification of hyponymy-hyperonymy relations. En *Modern Computational Models of Semantic Discovery in Natural Language*, págs. 48–70. IGI Global (2015)

Al-Yahya, M., Al-Malak, S., y Aldhubayi, L. Ontological lexicon enrichment: The BA-DEA system for semi-automated extraction of antonymy relations from Arabic language corpora. *Malaysian Journal of Computer Science* **29**(1):56–73 (2016)

Alarcón, R., Bach, C., y Sierra, G. Extracción de contextos definitorios en corpus especializados: Hacia la elaboración de una herramienta de ayuda terminográfica. *Revista española de lingüística* **37**(1):215–246 (2007)

Álvaro, J.F.V. Las clases de palabras y sus accidentes en la"Gramática general"de Gómez Hermosilla. *Anuario de Letras: Lingüística y filología* (21):5–45 (1983)

Angeli, G., Premkumar, M.J., y Manning, C.D. Leveraging linguistic structure for open domain information extraction. En *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)* (2015)

Arroyo-Fernández, I., Torres-Moreno, J.M., Sierra, G., y Adrián, L. Automatic Text Summarization by Non-Topic Relevance Estimation. En *8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KDIR 2017)*. Madeira, Portugal (2017)

Arroyo-Fernández, I., Méndez-Cruz, C.F., Sierra, G., Torres-Moreno, J.M., y Sidorov, G. Unsupervised sentence representations as word information series: Revisiting TF–IDF. *Computer Speech & Language* **56**:107–129 (2019)

Baroni, M. y Lenci, A. How we BLESSed distributional semantic evaluation. En *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, págs. 1–10. Association for Computational Linguistics (2011)

Baroni, M., Dinu, G., y Kruszewski, G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. En *ACL (1)*, págs. 238–247 (2014)

Başkaya, O., Sert, E., Cirik, V., y Yuret, D. Ai-ku: Using substitute vectors and co-occurrence modeling for word sense induction and disambiguation. En *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, págs. 300–306 (2013)

Bel-Enguix, G., Gómez-Adorno, H., Reyes-Magaña, J., y Sierra, G. Wan2vec: Embeddings learned on word association norms. *Semantic Web* **10**(6):991–1006 (2019)

Bel-Enguix, G., Gómez-Adorno, H., Pimentel, A., Ojeda-Trueba, S.L., y Aguilar-Vizuet, B. Negation Detection on Mexican Spanish Tweets: The T-MexNeg Corpus. *Applied Sciences* **11**(9):3880 (2021)

Bentivogli, L., Dagan, I., y Magnini, B. The recognizing textual entailment challenges: Datasets and methodologies. En *Handbook of Linguistic Annotation*, págs. 1119–1147. Springer (2017)

Biemann, C. Ontology learning from text: A survey of methods. En *LDV forum*, tomo 20, págs. 75–93 (2005)

Biemann, C., Faralli, S., Panchenko, A., y Ponzetto, S.P. A framework for enriching lexical semantic resources with distributional semantics. *Natural Language Engineering* **24**(2):265–312 (2018)

Blanco, E. y Moldovan, D. Semantic representation of negation using focus detection. En *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, págs. 581–589 (2011)

Bojanowski, P., Grave, E., Joulin, A., y Mikolov, T. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics* **5**:135–146 (2017)

Brewster, C., Alani, H., Dasmahapatra, S., y Wilks, Y. Data driven ontology evaluation (2004)

Bullinaria, J.A. y Levy, J.P. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods* **39**(3):510–526 (2007)

Bullinaria, J.A. y Levy, J.P. Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and SVD. *Behavior research methods* **44**(3):890–907 (2012)

Burton, G.O. y Rhetoricae, S. Figures of repetition. *Silva Rhetoricae: The Forest of Rhetoric* (2007)

Cardellino, C. Spanish Billion Words Corpus and Embeddings (2019)

Castillo, G. y Sierra, G. Algoritmo flexibilizado de agrupamiento semántico. *Estudios de Lingüística Aplicada* **21**:38 (2003)

Castro Rolón, B., Sierra, G., Torres-Moreno, J.M., y Da Cunha, I. El discurso y la sem ántica como recursos para la detecci ón de similitud textual. En *Proceedings of the III RST Meeting (8th Brazilian Symposium in Information and Human Language Technology, STIL 2011). Brazilian Computer Society, Cuiabá, Brasil* (2011)

Charles, W.G. y Miller, G.A. Contexts of antonymous adjectives. *Applied psycholinguistics* **10**(3):357–375 (1989)

Chen, D. y Manning, C.D. A Fast and Accurate Dependency Parser using Neural Networks. En *EMNLP*, págs. 740–750 (2014)

Chen, X., Liu, Z., y Sun, M. A Unified Model for Word Sense Representation and Disambiguation. En *EMNLP*, págs. 1025–1035 (2014)

Chen, Y. y Skiena, S. Building sentiment lexicons for all major languages. En *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, págs. 383–389 (2014)

Chen, Y., Perozzi, B., Al-Rfou, R., y Skiena, S. The expressive power of word embeddings. *arXiv preprint arXiv:1301.3226* (2013)

Church, K. y Hanks, P. Word association norms, mutual information, and lexicography. *Computational linguistics* **16**(1):22–29 (1990)

Conde, X.F. An introduction to syntax according to Generative Grammar Theories. *Romania Minor* (2005)

Coşeriu, E. *Gramatica, semantica, universales: estudios de linguistica funcional*. Editorial Gredos (1987)

Cruse, D.A., Cruse, D.A., Cruse, D.A., y Cruse, D.A. *Lexical semantics*. Cambridge university press (1986)

Cruz Domínguez, I. *El sintagma nominal en la extracción de relaciones léxico-semánticas de contextos definitorios: el caso de la preposición "de".* Tesis Doctoral, Universidad Nacional Autónoma de México (2011)

Dagan, I., Glickman, O., y Magnini, B. The pascal recognising textual entailment challenge. En *Machine Learning Challenges Workshop*, págs. 177–190. Springer (2005)

Damerau, F.J. The Use of Function Word Frequencies as Indicators of Style. *Computers and the Humanities* **9**(6):271–280 (1975)

Davidov, D. y Rappoport, A. Efficient unsupervised discovery of word categories using symmetric patterns and high frequency words. En *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, págs. 297–304 (2006)

Davidov, D., Tsur, O., y Rappoport, A. Enhanced sentiment learning using twitter hashtags and smileys. En *Proceedings of the 23rd international conference on computational linguistics: posters*, págs. 241–249. Association for Computational Linguistics (2010a)

Davidov, D., Tsur, O., y Rappoport, A. Semi-supervised recognition of sarcastic sentences in twitter and amazon. En *Proceedings of the fourteenth conference on computational natural language learning*, págs. 107–116. Association for Computational Linguistics (2010b)

De Boom, C., Van Canneyt, S., Demeester, T., y Dhoedt, B. Learning representations for tweets through word embeddings. En *Benelearn* (2016)

Devlin, J., Chang, M.W., Lee, K., y Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018)

Dorantes, M.A., Pimentel, A., Sierra, G., Bel-Enguix, G., y Molina, C. Extracción automática de definiciones analíticas y relaciones semánticas de hiponimia-hiperonimia con un sistema basado en patrones lingüísticos. *Linguamática* **9**(2):33–44 (2017)

Drymonas, E., Zervanou, K., y Petrakis, E.G. Unsupervised ontology acquisition from plain texts: the OntoGain system. En *International Conference on Application of Natural Language to Information Systems*, págs. 277–287 (2010)

Dubremetz, M. *Detecting Rhetorical Figures based on repetition of words: Chiasmus, epanaphora, epiphora*. Tesis Doctoral, Acta Universitatis Upsaliensis (2017)

Edmonds, P. y Hirst, G. Near-synonymy and lexical choice. *Computational linguistics* **28**(2):105–144 (2002)

Erk, K. y Padó, S. A structured vector space model for word meaning in context. En *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, págs. 897–906 (2008)

Española, R.A. Nueva gramática de la lengua española (2009)

Española, R.A. y Madrid, E. *Diccionario de la lengua española*, tomo 22. Real academia española Madrid (2001)

Fancellu, F., Lopez, A., y Webber, B. Neural networks for negation scope detection. En *Proceedings of the 54th annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, págs. 495–504 (2016)

Faruqui, M., Dodge, J., Jauhar, S.K., Dyer, C., Hovy, E., y Smith, N.A. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166* (2014)

Fellbaum, C. Co-occurrence and antonymy. *International journal of lexicography* **8**(4):281–303 (1995)

Ferrero, J., Agnes, F., Besacier, L., y Schwab, D. Using Word Embedding for Cross-Language Plagiarism Detection. *arXiv preprint arXiv:1702.03082* (2017)

Finkel, J.R., Grenager, T., y Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. En *Proceedings of the 43rd annual meeting on association for computational linguistics*, págs. 363–370 (2005)

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., y Ruppin, E. Placing search in context: The concept revisited. En *Proceedings of the 10th international conference on World Wide Web*, págs. 406–414 (2001)

Firth, J.R. A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis* (1957)

Fung, P. y Church, K.W. K-vec: A new approach for aligning parallel texts. En *Proceedings of the 15th conference on Computational linguistics-Volume 2*, págs. 1096–1102 (1994)

Gagné, A.M. y L'Homme, M.C. Opposite relationships in terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* **22**(1):30–51 (2016)

Ganuza, P.R. La delimitación de las unidades fraseológicas (UF) en la investigación alemana y española. *Interlingüística* (17):905–914 (2006)

Goldsmith, J.A. The Legacy of Zellig Harris: Language and Information into the 21st Century, vol. 1: Philosophy of Science, Syntax and Semantics. *Language* **81**(3):719–736 (2005)

Gómez-Adorno, H., Reyes-Magaña, J., Bel-Enguix, G., y Sierra, G. Spanish Word Embeddings Learned on Word Association Norms. En *AMW* (2019)

Goyal, T. y Durrett, G. Evaluating Factuality in Generation with Dependency-level Entailment. *arXiv preprint arXiv:2010.05478* (2020)

Grefenstette, G. Finding semantic similarity in raw text: The Deese antonyms. En *Fall Symposium Series, Working Notes, Probabilistic Approaches to Natural Language*, págs. 61–65 (1992)

Gruber, T.R. A translation approach to portable ontology specifications. *Knowledge acquisition* **5**(2):199–220 (1993)

Guo, J., Che, W., Wang, H., y Liu, T. Revisiting Embedding Features for Simple Semi-supervised Learning. En *EMNLP*, págs. 110–120 (2014)

Hallebeek, J. Las palabras funcionales del español. *Boletín de la* (1986)

Harris, Z.S. Distributional Structure. *Word* **10**(2-3):146–162 (1954). https://doi.org/10.1080/00437956.1954.11659520

Harris, Z.S. Co-Occurrence and Transformation in Linguistic Structure. *Language* **33**(3):283–340 (1957)

Hassan, S. y Mihalcea, R. Cross-lingual semantic relatedness using encyclopedic knowledge. En *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, págs. 1192–1201 (2009)

Hazem, A. y Daille, B. Word embedding approach for synonym extraction of multi-word terms. En *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018)

He, H., Gimpel, K., y Lin, J.J. Multi-Perspective Sentence Similarity Modeling with Convolutional Neural Networks. En *EMNLP*, págs. 1576–1586 (2015)

Hearst, M.A. Automatic acquisition of hyponyms from large text corpora. En *Proceedings of the 14th conference on Computational linguistics-Volume 2*, págs. 539–545 (1992)

Henriksson, A., Moen, H., Skeppstedt, M., Eklund, A.M., Daudaravicius, V., y Hassel, M. Synonym extraction of medical terms from clinical text using combinations of word space models. *Proceedings of Semantic Mining in Biomedicine (SMBM 2012)* págs. 10–17 (2012)

Henry, S. y Sands, A. VRep at SemEval-2016 Task 1 and Task 2: A System for Interpretable Semantic Similarity. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, págs. 577–583 (2016)

Hill, F., Cho, K., Jean, S., Devin, C., y Bengio, Y. Embedding word similarity with neural machine translation. *arXiv preprint arXiv:1412.6448* (2014)

Hill, F., Reichart, R., y Korhonen, A. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* **41**(4):665–695 (2015)

Honnibal, M. y Montani, I. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017). To appear

Horn, L.R. y Wansing, H. Negation.–Edward N. Zalta (toim.) (2017)

Iacobacci, I., Pilehvar, M.T., y Navigli, R. SensEmbed: Learning Sense Embeddings for Word and Relational Similarity. En *ACL (1)*, págs. 95–105 (2015)

Iliná, N. La fraseología española contemporánea: estado de la cuestión. En *Actas de la II Conferencia de hispanistas de Rusia* (2000)

Ireland, M.E., Slatcher, R.B., Eastwick, P.W., Scissors, L.E., Finkel, E.J., y Pennebaker, J.W. Language Style Matching Predicts Relationship Initiation and Stability. *Psychological Science* **22**(1):39–44 (2011)

Ismail, R., Rahman, N.A., y Bakar, Z.A. Extractions of Synonym Relations from English Translated Quran Using Seed Word Patterns (2017)

Jana, A. y Goyal, P. Network Features Based Co-hyponymy Detection. *arXiv preprint arXiv:1802.04609* (2018)

Jiménez-Zafra, S.M., Taulé, M., Martín-Valdivia, M.T., Urena-López, L.A., y Martí, M.A. SFU Review SP-NEG: a Spanish corpus annotated with negation for sentiment analysis. a typology of negation patterns. *Language Resources and Evaluation* **52**(2):533–569 (2018)

Jiménez-Zafra, S.M., Morante, R., Blanco, E., Valdivia, M.T.M., y Lopez, L.A.U. Detecting negation cues and scopes in Spanish. En *Proceedings of The 12th Language Resources and Evaluation Conference*, págs. 6902–6911 (2020)

Johansson, R. y Pina, L.N. Embedding a Semantic Network in a Word Space. En *HLT-NAACL*, págs. 1428–1433 (2015)

Jolly, H.B. Teaching Basic Function Words. *The Reading Teacher* **35**(2):136–140 (1981)

Jones, S. *Antonymy: A corpus-based perspective*. Routledge (2003)

Jones, S., Paradis, C., Murphy, M.L., y Willners, C. Googling for 'opposites': A Web-based study of antonym canonicity. *Corpora* **2**(2):129–155 (2007)

Justeson, J.S. y Katz, S.M. Co-occurrences of antonymous adjectives and their contexts. *Computational linguistics* **17**(1):1–19 (1991)

Kader, M.A., Boedihardjo, A.P., Naim, S.M., y Hossain, M.S. Contextual Embedding for Distributed Representations of Entities in a Text Corpus. En *Proceedings of the 5th International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, págs. 35–50 (2016)

Kiela, D., Hill, F., y Clark, S. Specializing word embeddings for similarity or relatedness. En *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, págs. 2044–2048 (2015)

Kim, J., Rousseau, F., y Vazirgiannis, M. Convolutional Sentence Kernel from Word Embeddings for Short Text Categorization. En *EMNLP*, págs. 775–780 (2015)

Komachi, T.K..M. Building a Monolingual Parallel Corpus for Text Simplification Using Sentence Similarity Based on Alignment between Word Embeddings (2016)

Kovatchev, V., Gold, D., Martí, M.A., Salamó, M., y Zesch, T. Decomposing and Comparing Meaning Relations: Paraphrasing, Textual Entailment, Contradiction, and Specificity. En *Proceedings of The 12th Language Resources and Evaluation Conference*, págs. 5782–5791 (2020)

Kozareva, Z., Riloff, E., y Hovy, E. Semantic class learning from the web with hyponym pattern linkage graphs. *Proceedings of ACL-08: HLT* págs. 1048–1056 (2008)

Kozima, H. Text segmentation based on similarity between words. En *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, págs. 286–288 (1993)

Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q. *et al.* From Word Embeddings To Document Distances. En *ICML*, tomo 15, págs. 957–966 (2015)

Larreta Zulategui, J.P. En torno a la semántica de las colocaciones fraseológicas. *ELUA. Estudios de Lingüística, N. 16 (2002); pp. 121-138* (2002)

Le, Q.V. y Mikolov, T. Distributed Representations of Sentences and Documents. En *ICML*, tomo 14, págs. 1188–1196 (2014)

Lenci, A. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics* **20**(1):1–31 (2008)

Lev, G., Klein, B., y Wolf, L. In defense of word embedding for generic text representation. En *International Conference on Applications of Natural Language to Information Systems*, págs. 35–50 (2015)

Levy, O. y Goldberg, Y. Dependency-based word embeddings. En *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, tomo 2, págs. 302–308 (2014a)

Levy, O. y Goldberg, Y. Neural word embedding as implicit matrix factorization. En *Advances in neural information processing systems*, págs. 2177–2185 (2014b)

Levy, O., Remus, S., Biemann, C., y Dagan, I. Do supervised distributional methods really learn lexical inference relations? En *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, págs. 970–976 (2015)

Liebeck, M., Pollack, P., Modaresi, P., y Conrad, S. Hhu at semeval-2016 task 1: Multiple approaches to measuring semantic textual similarity. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, págs. 595–601 (2016)

Liu, B. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* **5**(1):1–167 (2012)

Lobanova, A., Van der Kleij, T., y Spenader, J. Defining antonymy: A corpus-based study of opposites by lexico-syntactic patterns. *International Journal of Lexicography* **23**(1):19–53 (2010)

Lowe, D.G. Object recognition from local scale-invariant features. En *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, tomo 2, págs. 1150–1157 (1999)

Lyons, J. *Semantics: Volume 2*, tomo 2. Cambridge University Press (1977)

Marcus, M. New Trends in Natural Language Processing: Statistical Natural Language Processing. *Proceedings of the National Academy of Sciences of the United States of America* **92**(22):10052–10059 (1995)

Marton, Y., El Kholy, A., y Habash, N. Filtering antonymous, trend-contrasting, and polarity-dissimilar distributional paraphrases for improving statistical machine translation. En *Proceedings of the Sixth Workshop on Statistical Machine Translation*, págs. 237–249 (2011)

Melamud, O., Dagan, I., Goldberger, J., Szpektor, I., y Yuret, D. Probabilistic modeling of joint-context in distributional similarity. En *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, págs. 181–190 (2014)

Melamud, O., Dagan, I., y Goldberger, J. Modeling word meaning in context with substitute vectors. En *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, págs. 472–482 (2015)

Melamud, O., McClosky, D., Patwardhan, S., y Bansal, M. The role of context types and dimensionality in learning word embeddings. *arXiv preprint arXiv:1601.00893* (2016)

Mettinger, A. *et al. Aspects of semantic opposition in English*. Oxford University Press (1994)

Mijangos, V., Sierra, G., y Montes, A. Sentence level matrix representation for document spectral clustering. *Pattern Recognition Letters* **85**:29–34 (2017)

Mijangos de la Cruz, V.G. Agrupamiento temático de documentos basado en un modelo de similitud textual (2015)

Mikołajczak-Matyja, N. *et al.* The Prototypicality of Semantic Opposition in the Light of Linguistic Studies and Psycholinguistic Experiments. *Studies in Polish Linguistics* **13**(1):1–23 (2018)

Mikolov, T., Chen, K., Corrado, G., y Dean, J. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013a)

Mikolov, T., Le, Q.V., y Sutskever, I. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* (2013b)

Mikolov, T., Yih, W.t., y Zweig, G. Linguistic regularities in continuous space word representations. En *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, págs. 746–751 (2013c)

Miller, G.A. WordNet: a lexical database for English. *Communications of the ACM* **38**(11):39–41 (1995)

Miller, G.A. *WordNet: An electronic lexical database*. MIT press (1998)

Miller, G.A. y Charles, W.G. Contextual correlates of semantic similarity. *Language and cognitive processes* **6**(1):1–28 (1991)

Mnih, A. y Hinton, G.E. A scalable hierarchical distributed language model. En *Advances in neural information processing systems*, págs. 1081–1088 (2009)

Morin, F. y Bengio, Y. Hierarchical Probabilistic Neural Network Language Model. En *Aistats*, tomo 5, págs. 246–252 (2005)

Mousavi, H., Kerr, D., Iseli, M., y Zaniolo, C. Ontoharvester: An unsupervised ontology generator from free text. *UCLA* (2013)

Mrkšić, N., Séaghdha, D.O., Thomson, B., Gašić, M., Rojas-Barahona, L., Su, P.H., Vandyke, D., Wen, T.H., y Young, S. Counter-fitting word vectors to linguistic constraints. *arXiv preprint arXiv:1603.00892* (2016)

Na, J.C., Khoo, C., y Wu, P.H.J. Use of negation phrases in automatic sentiment classification of product reviews. *Library Collections, Acquisitions, and Technical Services* **29**(2):180–191 (2005)

Nivre, J. Incrementality in deterministic dependency parsing. En *Proceedings of the Workshop on Incremental Parsing: Bringing Engineering and Cognition Together*, págs. 50–57. Association for Computational Linguistics (2004)

Nugumanova, A., Baiburin, Y., Apayev, K., Tlebaldinova, A., y Rakhadilov, B. Automatic synonym and quasi-synonym extraction from user reviews on mobile devices. En *2019 19th International Conference on Control, Automation and Systems (ICCAS)*, págs. 1503–1507. IEEE (2019)

Orešković, M. *An Online Syntactic and Semantic Framework for Lexical Relations Extraction Using Natural Language Deterministic Model.* Tesis Doctoral, University of Zagreb. Faculty of Organization and Informatics. (2019)

Padó, S. y Lapata, M. Constructing semantic space models from parsed corpora. En *Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, págs. 128–135 (2003)

Pavlick, E., Rastogi, P., Ganitkevitch, J., Van Durme, B., y Callison-Burch, C. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. En *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, págs. 425–430 (2015)

Pekar, V. Discovery of event entailment knowledge from text corpora. *Computer Speech & Language* **22**(1):1–16 (2008)

Pennington, J., Socher, R., y Manning, C.D. Glove: Global Vectors for Word Representation. En *EMNLP*, tomo 14, págs. 1532–1543 (2014)

Piantadosi, S.T. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review* **21**(5):1112–1130 (2014)

Pustejovsky, J. The generative lexicon. *Computational linguistics* **17**(4):409–441 (1991)

Reinberger, M.L. y Spyns, P. Unsupervised text mining for the learning of dogma-inspired ontologies. *Ontology Learning from Text: Methods, Applications and Evaluation* **123**:29–43 (2005)

Riedl, M. y Biemann, C. Scaling to large3 data: An efficient and effective method to compute distributional thesauri. En *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, págs. 884–890 (2013)

Roget, P.M. *Roget's Thesaurus of English Words and Phrases...* TY Crowell Company (1911)

Roller, S. y Erk, K. Relations such as hypernymy: Identifying and exploiting hearst patterns in distributional vectors for lexical entailment. *arXiv preprint arXiv:1605.05433* (2016)

Roller, S., Kiela, D., y Nickel, M. Hearst patterns revisited: Automatic hypernym detection from large text corpora. *arXiv preprint arXiv:1806.03191* (2018)

Roth, M. y Im Walde, S.S. Combining word patterns and discourse markers for paradigmatic relation classification. En *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, págs. 524–530 (2014)

Roth, M. y Upadhyay, S. Combining Discourse Markers and Cross-lingual Embeddings for Synonym–Antonym Classification. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, págs. 3899–3905 (2019)

Ruder, S. On word embeddings - Part 1. http://ruder.io/word-embeddings-1/ (2016a)

Ruder, S. On word embeddings - Part 2: Approximating the Softmax. http://ruder.io/word-embeddings-softmax (2016b)

Ruder, S. On word embeddings - Part 3: The secret ingredients of word2vec. http://ruder.io/secret-word2vec/ (2016c)

Saha, T.K., Joty, S., Hassan, N., y Hasan, M.A. Dis-S2V: Discourse Informed Sen2Vec. *arXiv preprint arXiv:1610.08078* (2016)

Sarkar, S., Das, D., Pakray, P., y Gelbukh, A. JUNITMZ at SemEval-2016 task 1: Identifying semantic similarity using Levenshtein ratio. En *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, págs. 702–705 (2016)

Sasaki, Y. *et al.* The truth of the f-measure. 2007 (2007)

Schwartz, R., Tsur, O., Rappoport, A., y Koppel, M. Authorship attribution of micro-messages. En *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, págs. 1880–1891 (2013)

Schwartz, R., Reichart, R., y Rappoport, A. Minimally supervised classification to semantic categories using automatically acquired symmetric patterns. En *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, págs. 1612–1623 (2014)

Schwartz, R., Reichart, R., y Rappoport, A. Symmetric pattern based word embeddings for improved word similarity prediction. En *Proceedings of the nineteenth conference on computational natural language learning*, págs. 258–267 (2015)

Schwartz, R., Reichart, R., y Rappoport, A. Symmetric patterns and coordinations: Fast and enhanced representations of verbs and adjectives. En *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, págs. 499–505 (2016)

Shah, V. y Rekh, P. A survey: Importance of negation in sentiment analysis. *International Journal of Emerging Technology and Advanced Engineering* **4**(3):70–73 (2014)

Shi, J. y Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence* **22**(8):888–905 (2000)

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., y Chanona-Hernández, L. Syntactic dependency-based n-grams as classification features. En *Mexican International Conference on Artificial Intelligence*, págs. 1–11 (2012)

Sidorov, G., Velasquez, F., Stamatatos, E., Gelbukh, A., y Chanona-Hernández, L. Syntactic n-grams as machine learning features for natural language processing. *Expert Systems with Applications* **41**(3):853–860 (2014)

Sierra, G. y Alarcón, R. Identification of recurrent patterns to extract definitory contexts. En *International Conference on Intelligent Text Processing and Computational Linguistics*, págs. 436–438. Springer (2002)

Snow, R., Jurafsky, D., y Ng, A.Y. Learning syntactic patterns for automatic hypernym discovery. En *Advances in neural information processing systems*, págs. 1297–1304 (2005)

Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C. *et al.* Recursive deep models for semantic compositionality over a sentiment treebank. En *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, tomo 1631, pág. 1642 (2013)

Torres-Moreno, J.M., Molina, A., y Sierra, G. La energía textual como medida de distancia en agrupamiento de definiciones. En *International Conference on Statistical Analysis of Textual Data* (2010)

Turian, J., Ratinov, L., y Bengio, Y. Word representations: a simple and general method for semi-supervised learning. En *Proceedings of the 48th annual meeting of the association for computational linguistics*, págs. 384–394 (2010)

Vidal, E. y Leonetti, M. Categorías funcionales y semántica procedimental. *Cien años de investigación semántica: de Michel Bréal a la actualidad* **1**:363–378 (1997)

Víta, M. From Building Corpora for Recognizing Faceted Entailment to Recognizing Relational Entailment. En *FedCSIS (Position Papers)*, págs. 33–38 (2018)

Wang, T. y Hirst, G. Exploring patterns in dictionary definitions for synonym extraction. *Natural Language Engineering* **18**(3):313–342 (2012)

Wang, W., Thomas, C., Sheth, A., y Chan, V. Pattern-based synonym and antonym extraction. En *Proceedings of the 48th annual southeast regional conference*, págs. 1–4 (2010)

Weeds, J., Clarke, D., Reffin, J., Weir, D., y Keller, B. Learning to distinguish hypernyms and co-hyponyms. En *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, págs. 2249–2259. Dublin City University and Association for Computational Linguistics (2014)

Widdows, D. y Dorow, B. A graph model for unsupervised lexical acquisition. En *Proceedings of the 19th international conference on Computational linguistics-Volume 1*, págs. 1–7. Association for Computational Linguistics (2002)

Yatbaz, M.A., Sert, E., y Yuret, D. Learning syntactic categories using paradigmatic representations of word context. En *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, págs. 940–951 (2012)

Yule, G. *The study of language*. Cambridge university press (2020)

Zanzotto, F.M. y Pennacchiotti, M. Expanding textual entailment corpora fromwikipedia using co-training. En *Proceedings of the 2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, págs. 28–36 (2010)

Zhang, L., Li, J., y Wang, C. Automatic synonym extraction using Word2Vec and spectral clustering. En *2017 36th Chinese Control Conference (CCC)*, págs. 5629–5632. IEEE (2017)