Benemérita Universidad Autónoma de Puebla

Facultad de Ciencias de la Computación

Evaluación de conceptos a través de medidas de similitud semántica en taxonomías de dominio restringido

Tesis que para obtener el título de:
Licenciatura en Ingeniería en Ciencias de la Computación

Presenta:
Aimee Cecilia Hernández García

Asesores:
Dra. Mireya Tovar Vidal
Dr. José de Jesús Lavalle Martínez

Febrero de 2019

Dedicatoria

La presente tesis se la dedico principalmente a mis padres que con su apoyo y amor incondicional he logrado formarme tanto personal como profesionalmente. A mi amigo Pedro por siempre motivarme y estar en los buenos y malos momentos. A mi amigo Enrique por el apoyo durante toda la carrera. A mis amigos Abdiel y Areli por los buenos momentos que hemos pasado juntos.

Agradecimientos

Quiero agradecer al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca con número de registro 27483 recibida durante el desarrollo de esta tesis en el periodo Marzo 2018 - Enero 2019 mediante el proyecto número 257357 de Ciencia Básica en la modalidad de Joven Investigador, titulado "Estimulación de parámetros para métricas de evaluación de ontologías", en el área IV de Humanidades y Ciencia de la Conducta.

Resumen

La similitud semántica, medida que se utiliza para saber el grado de similitud entre dos clases o conceptos en una ontología, es de gran importancia debido a sus extensas aplicaciones en campos como biomedicina, donde se han desarrollado ontologías para comparar genes, proteínas, compuestos químicos, entidades anatómicas y enfermedades, en la geoinformática ha servido para encontrar similitud en características geográficas, en lingüística computacional y procesamiento de lenguaje natural donde se utiliza para el análisis de sentimientos, el entendimiento del lenguaje natural, la traducción automática y la web semántica. En la actualidad existe una extensa variedad de medidas de similitud semántica para ontologías, que se clasifican generalmente en dos grandes grupos: las medidas basadas en conocimiento y las medidas basadas en corpus. En esta investigación se pretende implementar algunas de las medidas de similitud basadas en conocimiento que han sido propuestas en la literatura y que se consideren candidatas para evaluar su comportamiento empleando diferentes juicios de evaluación y al final hacer una comparativa y concluir con ventajas y desventajas de las medidas de similitud.

Índice general

Dedicatoria I

Agradecimientos II

Resumen III

Índice de figuras VI

Índice de tablas VII

Índice de algoritmos X

1. Introducción 1

1.1. Planteamiento del problema 2

1.2. Objetivos 2

1.2.1. Objetivo general 3

1.2.2. Objetivos específicos 3

1.3. Antecedentes 3

1.4. Justificación 5

1.5. Metodología 5

1.6. Estructura del documento de tesis 6

2. Estado del arte 7

2.1. Medidas de similitud semántica 7

2.1.1. Medidas de similitud semántica basadas en estructura 8

2.1.2. Medidas de similitud semántica basadas en características . . 9

2.1.3. Medidas de similitud semántica basadas en contenido de información . . . 10
2.1.4. Evaluación de ontologías . . . 11

3. Marco teórico **14**
3.1. Procesamiento de Lenguaje Natural . . . 14
3.1.1. Niveles del procesamiento de lenguaje natural . . . 15
3.1.2. Aplicaciones del procesamiento de lenguaje natural . . . 16
3.2. Ontologías . . . 17
3.2.1. Principales componentes de una ontología . . . 17
3.2.2. Clasificación de ontologías . . . 18
3.3. Ingeniería ontológica . . . 19
3.4. Semántica . . . 24
3.4.1. Relaciones semánticas . . . 24
3.5. Similitud semántica . . . 27
3.5.1. Medidas de similitud semántica . . . 28
3.5.2. Clasificación de las medidas de similitud semántica . . . 28
3.5.3. Medidas basadas en conocimiento . . . 28
3.6. Herramientas para ontologías . . . 36
3.6.1. XML (Extensible Markup Language) . . . 36
3.6.2. RDF (Resource Description Framework) . . . 36
3.6.3. OWL (Web Ontology Language) . . . 37
3.6.4. Prótegé . . . 38
3.6.5. Sematch . . . 39
3.6.6. SML (Semantic Measures Library) . . . 40

4. Diseño de la propuesta de solución **43**
4.1. Algoritmo propuesto . . . 43
4.1.1. Preprocesamiento . . . 46
4.1.2. Cálculo de la similitud semántica . . . 46
4.1.3. Cálculo de los umbrales . . . 46
4.1.4. Evaluación . . . 47
4.1.5. Exactitud . . . 49

5. Resultados experimentales **51**
5.1. Conjuntos de datos 51
5.1.1. Ontologías con anotación de expertos 51
5.1.2. Ontologías del dominio de biomedicina 52
5.1.3. Ontologías del dominio de sensores 54
5.2. Evaluación con anotaciones de expertos 56
5.2.1. Resultados de las medidas basadas en estructura 56
5.2.2. Resultados de las medidas basadas en características 59
5.2.3. Resultados de las medidas basadas en contenido de información 62
5.2.4. Resultados de las medidas híbridas 66
5.2.5. Resultados en todas las categorías de las medidas 69
5.2.6. Discusión 70
5.3. Evaluación de ontologías de dominio de biomedicina 71
5.3.1. Resultados de las medidas basadas en estructura 72
5.3.2. Resultados de las medidas basadas en características 74
5.3.3. Resultados de las medidas basadas en contenido de información 75
5.3.4. Resultados de las medidas híbridas 77
5.3.5. Resultados en todas las categorías de las medidas 78
5.3.6. Discusión 80
5.4. Resultados con el conjunto de datos en el dominio de sensores 81
5.4.1. Resultados de las medidas basadas en estructura 81
5.4.2. Resultados de las medidas basadas en características 82
5.4.3. Resultados de las medidas basadas en contenido de información 83
5.4.4. Resultados de las medidas híbridas 84
5.4.5. Resultados en todas las categorías de medidas 85
5.4.6. Discusión 86
5.5. Discusión general 86

Conclusiones **88**

Bibliografía **90**

Índice de figuras

3.1. Componentes en una ontología. 18
3.2. Tipos de ontologías y sus relaciones [17]. 19
3.3. Representación de la relación entre hiperónimo-hipónimo. 25
3.4. Ejemplo de una taxonomía 27
3.5. Anatomía de un documento XML [13] 37
3.6. Ejemplo de un documento RDF Schema [13] 38
3.7. Estructura de un documento OWL [13] 39
3.8. Interfaz de Prótegé versión escritorio [28] 40
3.9. Interfaz de Prótegé versión web [28] 41
3.10. Demostración web del *framework* Sematch [64][65] 42

4.1. Algoritmo para evaluar las relaciones taxonómicas 45

5.1. Camino en la ontología SCORM que causa conflicto. 65

Índice de tablas

2.1. Tabla comparativa de las medidas de similitud semánticas 12

5.1. Conjunto de datos con anotaciones de expertos 52
5.2. Total de relaciones taxonómicas revisadas por los expertos 52
5.3. Expertos para cada ontología 52
5.4. Conjunto de datos del dominio de biomedicina 54
5.5. Dominio de entrada y resultados de los sistemas de recomendación, para el dominio de biomedicina 54
5.6. Ontologías de sensores 56
5.7. Umbrales de las medidas basadas en estructura en la ontología IA . . 56
5.8. Exactitudes de las medidas basadas en estructura para la ontología IA 57
5.9. Umbrales de las medidas basadas en estructura para la ontología SCORM 57
5.10. Exactitudes de las medidas basadas en estructura para la ontología SCORM 58
5.11. Umbrales de las medidas basadas en estructura en la ontología EOR . 58
5.12. Exactitudes de las medidas basadas en estructura en la ontología EOR 59
5.13. Umbrales de las medidas basadas en características de la ontología IA 60
5.14. Exactitudes de las medidas basadas en características para la ontología IA 60
5.15. Umbrales de las medidas basadas en características de la ontología SCORM 61
5.16. Exactitudes de las medidas basadas en características para la ontología SCORM 61

5.17. Umbrales de las medida basadas en características de la ontología EOR 62
5.18. Exactitudes de las medidas basadas en estructura para la ontología EOR 62
5.19. Umbrales de las medidas basadas en contenido de información en la ontología IA 63
5.20. Exactitudes de las medidas basadas en contenido de información para la ontología IA 63
5.21. Umbrales de las medidas basadas en contenido de información en la ontología EOR 64
5.22. Exactitudes de las medidas basadas en contenido de información para la ontología EOR 65
5.23. Umbrales de cada medida híbrida en la ontología IA 66
5.24. Exactitudes de las medidas híbridas para la ontología IA 67
5.25. Umbrales de las medidas híbridas en la ontología EOR 68
5.26. Exactitudes de la ontología EOR con las medidas híbridas 68
5.27. Exactitud promedio para todas las categorías 69
5.28. Exactitud para todas las categorías 69
5.29. Ontologías sugeridas por cada uno de los sistemas de recomendación . 72
5.30. Umbrales de las medidas basadas en estructura en las ontologías de biomedicina 72
5.31. Exactitudes de las medidas basadas en estructura para las ontologías de biomedicina 73
5.32. Dominio de entrada y resultados de los sistemas 73
5.33. Umbrales de cada medida basadas en características en las ontologías de biomedicina 74
5.34. Exactitudes de las medidas basadas en características para las ontologías de biomedicina 74
5.35. Dominio de entrada y resultados de los sistemas 75
5.36. Umbrales de las medidas basadas en contenido de información en las ontologías de biomedicina 76
5.37. Exactitudes de las medidas basadas en contenido de información para las ontologías de biomedicina 76

5.38. Dominio de entrada y resultados de los sistemas 76
5.39. Umbrales de las medidas híbridas en las ontologías de biomedicina . . 77
5.40. Exactitudes de las medidas híbridas para las ontologías de biomedicina 78
5.41. Dominio de entrada y resultados de los sistemas 78
5.42. Exactitud para todas las categorías 79
5.43. Dominio de entrada y resultados de los sistemas 80
5.44. Umbrales de las medidas basadas en estructura en las ontologías de sensores 81
5.45. Exactitudes de las medidas basadas en estructura para las ontologías del dominio de sensores 81
5.46. Umbrales de las medidas basadas en características en las ontologías de sensores 82
5.47. Exactitudes de las medidas basadas en características para las ontologías del dominio de sensores 82
5.48. Umbral de las medidas basadas en contenido de información en las ontologías de sensores 83
5.49. Exactitud de las medidas basadas en contenido de información en las ontologías del dominio de sensores 83
5.50. Umbrales para las medidas híbridas 84
5.51. Exactitudes de las medidas híbridas para las ontologías del dominio de sensores 85
5.52. Exactitud para todas las categorías 85

Índice de algoritmos

1. Algoritmo propuesto 44
2. Función *compUmbral* 48
3. Función *compExperto* 48
4. Función *promedioSist* 50

Capítulo 1

Introducción

Desde su creación, la web se ha convertido en una herramienta de uso cotidiano para millones de personas en el mundo para el intercambio de información. Al pasar los años surgieron nuevas tecnologías que ayudaron a la web a convertise en dinámica y es entonces cuando nace el término de web 2.0, en donde esta web permite a las personas interactuar entre sí y compartir contenido. La facilidad y rapidez que ofrece la web 2.0 permite a los usuarios generar y compartir grandes cantidades de datos alrededor del mundo. Actualmente existen exuberantes cantidades de datos desorganizados y duplicados por toda la web debido a la mala estructuración que tienen, esto hace que sea difícil la tarea de consultar información en internet mediante un motor de búsqueda, lo que conlleva a no obtener los resultados esperados de la consulta hecha por el usuario. Ante esta problemática, Tim Berners-Lee [5] presenta la siguiente propuesta: La web semántica, en donde se presenta como una extensión de la web actual, que se le proporcionará de estructura para expresar el contenido de las páginas web de tal manera que las computadoras puedan procesarlas y así hacer posible la interacción entre computadoras así como entre computadoras y usuarios. Para que la web semántica funcione, toda la información debe estar semánticamente estructurada y marcada. Uno de los componentes de la web semántica son las ontologiás, el papel que toman las ontologías en la web semántica es que éstas se encargan de representar el conocimiento de la web para que pueda ser comprensible para las computadoras. Las ontologías definen, mediante una lógica formal, los conceptos de diferentes dominios del mundo real, además de definir las relaciones que

existen entre los conceptos del dominio. Las relaciones en las ontologías son vínculos concretos entre clases o instancias que llevan una semántica específica. Las medidas semánticas Este trabajo de investigación se enfocará en una parte de la validación de ontologías donde se busca evaluar los conceptos de una ontología de dominio con medidas de similitud semántica para decidir si la relación, en este caso de tipo "is-a" o taxonómica, hecha entre ellos es verdadera y bien concretada. Una relación taxonómica se toma verdadera cuando cumple con ciertos criterios como

1.1. Planteamiento del problema

Tradicionalmente la generación de ontologías de dominio se realiza de manera manual ya que se requiere de la participación de un ingeniero de conocimiento o de un ingeniero ontológico para su creación. Las aplicaciones actuales requieren de ontologías dinámicas y de gran extensión en sus conceptos que un ingeniero o varios no podrían manejar fácilmente, esto conlleva a tener un costo elevado en la realización de una ontología. Ante este problema, se ha introducido en los últimos años la generación (semi-) automática de ontologías o aprendizaje de ontologías. El aprendizaje de ontologías ayuda a la construcción de ontologías de forma semiautomática con supervisión del ingeniero de ontología. En el proceso de la generación semiautomática de ontologiás muy pocas son las propuestas que integran una evaluación automática en su proceso, por lo que la calidad de la ontología generada no puede ser comprobada.

En este trabajo, ante el problema de la calidad de las ontologías en su generación (semi-)automática, se investigaron medidas de similitud semántica para evaluar las relaciones de tipo "is-a" o taxonómicas en una ontología midiendo la similitud entre los conceptos que integran esta relación para conocer el grado de relación que existe entre ellos. Y así, dar un juicio automático de calidad para la ontología.

1.2. Objetivos

Los objetivos se dividen en dos partes, en el objetivo general se presenta el alcance de esta investigación, y en los objetivos principales se detallan las tareas que se

realizarán a lo largo de la investigación.

1.2.1. Objetivo general

El objetivo general de esta tesis es implementar medidas de similitud semántica basadas en conocimiento para la evaluación de conceptos en taxonomías de dominio restringido.

1.2.2. Objetivos específicos

En particular, los pasos a seguir con la finalidad de realizar esta tesis son:

- Implementar algunas de las medidas de similitud semántica que han sido propuestas.
- Evaluar las medidas de similitud y observar su desempeño.
- Realizar una comparativa entre las medidas de similitud implementadas.

1.3. Antecedentes

A continuación se presentan trabajos que son apoyo y antecedente para esta tesis:

En [52], se propuso un método para evaluar pares de palabras de entrada y medir el grado de similitud semántica. Este método no supervisado hizo uso de un vector prototipo calculado sobre la base de vectores representativos de pares de palabras que se construyen mediante el uso de fragmentos recogidos automáticamente de la red. En este artículo se consideraron dos clases semánticas: clase inclusión y parte-todo para determinar el grado de similitud semántica entre dos instancias de la ontología.

En [2], se desarrolló un modelo que resuelve el problema de similitud semántica en textos de diferente longitud creando un vector para cada texto y que posteriormente lo introducen en Weka para construir un modelo de clasificación basado en regresión logística. Los autores usan seis medidas mencionanadas a continuación: Leacock & Chodorow, Lesk, Wu & Palmer, Resnik, Lin y Jiang & Conrath. Este modelo fue planteado para resolver el problema de la tarea 3 del Semeval 2014. Se concluye que el modelo propuesto tiene un buen comportamiento en la detección del grado

de similitud semántica en párrafo-sentencia y sentencia-frase. Sin embargo, en frase-palabra y palabra-sentido no se obtuvieron buenos resultados.

En [1] se presentan dos modelos que fueron probados para inglés y español, donde su objetivo es detectar el grado de similitud semántica multilingüe, enfocado a pares de sentencias. El primer modelo se basa en aprendizaje supervisado y el segundo modelo en aprendizaje no supervisado. En el modelo supervisado, se obtiene un vector de características de cada par de sentencias para entrenar a un clasificador. Para asignar un grado de simiitud a las sentencias de prueba se utilizan las siguientes medidas: similitud coseno, similitud euclidiana, similitud Levenshtein y similitud Dice. En el modelo no supervisado, se calcula la similitud semántica de dos sentencias por medio de la intertextualidad, se reconstruye la sentencia de menor longitud por medio de la sentencia de mayor longitud. Mediante los resultados experimentales, se llega a la conclusión que en el modelo no supervisado se obtiene un mejor desempeño para ambos idiomas.

En [59], se exponen dos propuestas que se basan en similitud textual para determinar un juicio de las implicaciones textuales entre dos textos en dos idiomas diferentes: la primera propuesta utiliza la similitud léxica usando el coeficiente de Jaccard como medida de similitud y en la segunda propuesta hacen uso de la similitud semántica considerando los sinónimos de cada término de cada oración. Estos métodos fueron planteados para la tarea 8 para la evaluación de implicaciones textuales multilingüe para la sincronización de contenido del Semeval 2012. Con base en los resultados obtenidos en las dos propuestas se tiene que ambas muestran un comportamiento similar pero la propuesta semántica obtiene un mejor desempeño.

Basándonos en los antecedentes, el presente trabajo de investigación pretende implementar algunas de las medidas de similitud semántica basadas en conocimiento para ser aplicadas a conceptos de las relaciones taxonómicas de ontologías de dominio para que sean evaluadas midiendo el grado de similitud que tienen estos conceptos y posteriormente, realizar una comparación entre todas las medidas de similitud implementadas.

1.4. Justificación

La generación de ontologías de forma (semi-)automática es de gran ayuda para el ingeniero ontológico al agilizar el proceso de creación y diseño ya que a mayor tamaño de la ontología, la tarea se vuelve más difícil para un ser humano y se necesita de la ayuda de un proceso automatizado para reducir el costo en el tiempo y trabajo del ingeniero de ontologías. La evaluación en la generación (semi-)automática de las ontologías, es una tarea con la que se debe tener cuidado porque al ser un proceso automatizado y de reciente crecimiento, no se tiene un control de la calidad de la ontología creada.

En esta investigación se hará uso de las medidas de similitud semántica para hacer una evaluación a las relaciones taxonómicas de una ontología y emitir un juicio de calidad. Este trabajo aportará resultados de las pruebas experimentales realizadas para ampliar la investigación de la evaluación de ontologías centrado sólo en las relaciones taxonómicas.

1.5. Metodología

Con el fin de cumplir los objetivos se usará el modelo clásico para el desarrollo de software, el cual inicia con la obtención de requerimientos, seguida por el análisis, especificación, diseño, implementación y por último las pruebas.

Para tener un punto de referencia y base en el cual realizar el trabajo, se hará una búsqueda del estado del arte, que nos hará conocer y tener una visión de las propuestas que han sido planteadas durante los años, cuáles fueron sus bases, el problema que llevó a su desarrollo y los resultados que se obtuvieron. También, se hará una revisión general de los conceptos básicos de los fundamentos de web semántica, ontologías, similitud semántica y procesamiento de lenguaje natural.

Durante el desarrollo de la implementación, se usará la metodología de programación adecuada para implementar las medidas de similitud semántica y del sistema en general, lo que nos permitirá obtener resultados que nos servirán para someter a validación lo propuesto. Para realizar las pruebas se harán uso de ontologías de dominio donde se aplicarán juicios de evaluación. Primeramente se usarán ontologías que cuentan con una anotación de las relaciones taxonómicas realizadas por exper-

tos en el dominio, posteriormente se usarán ontologías en el dominio de biomedicina obtenidas desde NCBO BioPortal[1]. Otras ontologías usadas serán del dominio de sensores. Como siguiente paso se analizarán los resultados experimentales obtenidos de la evaluación de las relaciones taxonómicas de las ontologías. Finalmente, expondremos un análisis entre las medidas de similitud semántica, ventajas y desventajas para ayudar a investigadores a escoger la medida que se adecue a sus requerimientos para la validación de ontologías.

1.6. Estructura del documento de tesis

El presente trabajo de tesis se organiza de la siguiente manera:

- Capítulo 1, Introducción. Se muestra el planteamiento del problema, justificación, objetivos generales y específicos y por último la metodología.
- Capítulo 2, Estado del Arte. Se describen las contribuciones realizadas por diversos autores en el rubro de las medidas de similitud semántica y evaluación de ontologías.
- Capítulo 3, Marco teórico. Se exponen las bases teóricas de la investigación, útil para comprender el significado del contenido expuesto.
- Capítulo 4, Diseño. Se muestra el diseño del sistema propuesto para la resolución del problema planteado.
- Capítulo 5, Resultados. Se exponen a detalle los resultados obtenidos por el sistema propuesto.
- Capítulo 6, Conclusiones. En esta sección se muestran las conclusiones obtenidas al realizar este trabajo de tesis.

[1] https://bioportal.bioontology.org/

Capítulo 2

Estado del arte

A continuación se presentan trabajos relacionados con similitud semántica que servirán de base para el desarrollo de esta investigación. Particularmente estamos interesados en conocer los diferentes métodos de similitud semántica basada en ontologías.

2.1. Medidas de similitud semántica

Thabet Slimani [43] clasifica las medidas semánticas en tres clases o categorías y además presenta una evaluación basada en dos pruebas de rendimiento estándar. Las categorías son las siguientes: las medidas basadas en estructura; o las medidas de conteo de aristas, usan una función que calcula la medida de similitud semántica en la estructura jerárquica de una ontología, aquí se clasifica diferentes medidas que han sido propuestas tales como: el camino más corto, enlaces ponderados, medida de Hirst y St-onge(HSO), Wu & Palmer, Slimani, Li, y por ultimo Leacock & Chodorow. La segunda categoría son las medidas basadas en contenido de información, que como su nombre lo indica, usan el contenido de información de los conceptos para medir la similitud semántica entre dos conceptos o términos y se calcula en base a la frecuencia del término dentro de un corpus de texto dado. También se encuentran las medidas propuestas por Resnik, Lord, Lin, y Jiang & Conrath. La tercera categoría son las medidas basadas en características, que asumen que cada término está descrito por un conjunto de términos indicando sus propiedades o características, en esta categoría

se encuentran las medidas de Tversky, Similitud-X y Rodriguez. Las pruebas de rendimiento usadas fueron la de Miller & Charles y la de Rubenstein & Goodenough. Los resultados obtenidos son de ayuda para investigadores y profesionistas para que puedan elegir la medida que se ajusta a los requisitos de la aplicación a desarrollar.

2.1.1. Medidas de similitud semántica basadas en estructura

En [36] Rada propone una métrica llamada Distancia que evalúa el camino más corto o distancia semántica entre dos conceptos en una red semántica de relaciones jerárquicas, especialmente con relaciones "is-a", esta métrica tiene sus bases en la teoría de activación propagante. El autor afirma que mientras la red semántica sea mejor, las decisiones de similitud conceptual de la medida Distancia coinciden mejor con las decisiones de similitud conceptual de las personas.

En [62] se propone una esquema de representación que utiliza una medida de similitud semántica para resolver el problema de selección léxica en la traducción automática. Los autores definen la similitud de dos conceptos que se da por la cercanía de la relación en la jerarquía. Para el cálculo de la similitud de dos conceptos C_1 y C_2, se calcula el número de nodos que existen desde el concepto (C_1 o C_2) al superconcepto menos común, después se calcula el número de nodos del superconcepto menos común al concepto raíz. Los resultados experimentales muestran que al usar este esquema se puede corregir la selección léxica en coincidencias no exactas.

En [21] se propone una medida de similitud semántica para taxonomías de tipo "is-a" y ' 'has-a" que combina el camino más corto, la profundidad del subsumidor (subsumidor es el ancestro común más específico de dos conceptos en una ontología [37]), y la densidad semántica local. Donde el camino más corto y la profundidad del subsumidor se obtiene de una base de datos léxica, por otro lado, la densidad semántica local se obtiene desde el corpus. Los experimentos realizados mediante la prueba de rendimiento de Rubenstein-Goodenough indican que la mejor correlación en la literatura es de 0.8484 y la correlación de esta propuesta es de 0.8914. Esta medida supera las medidas publicadas anteriormente como las de Resnik y Jiang & Conrath y la propuesta de conteo de aristas.

2.1.2. Medidas de similitud semántica basadas en características

En [25] se propusieron métodos para medir la similitud en ontologías, donde se consideraron los niveles sintáctico y semántico. Durante el desarrollo se realizó una evaluación empírica de ontologías donde se hizo un estudio en el cual se usaron ontologías del dominio de turismo. Estas ontologías fueron creadas por estudiantes de licenciatura, se aplicaron las medidas propuestas para comparar las ontologías creadas por los estudiantes contra una ontología estándar de oro creada por un ingeniero ontológico experto, para que de esta manera se pueda encontar la mejor ontología que coincida con la ontología estándar de oro. Las medidas propuestas han resultado ser útiles para el mapeo entre dos ontologías y la integración de ontologías existentes en el proceso de ingeniería ontológica, así como facilitar la tarea de ingeniería ontológica colaborativa.

En [38] proponen un enfoque para medir la similitud de dos conceptos en diferentes ontologías utilizando la teoría de conjuntos. Los autores definen su medida de similitud propuesta basándose en la normalización del modelo de Tversky [53] y las funciones de intersección y diferencia de conjuntos. Este modelo de similitud detecta clases de entidad en ontologías mediante un proceso de coincidencia de cada componente de especificación en las representaciones de clases de entidad. Como este modelo detecta la mayoría de clases de entidad similares, es útil como un primer paso en la integración de la ontología.

En [47] se presenta una nueva medida que sigue el modelo propuesto por [53]. Esta medida se basa en la información taxonómica. Los autores consideran que un concepto se puede diferenciar semánticamente de otro al comparar el conjunto de conceptos que subsume a este concepto, es decir, el conjunto de conceptos ancestros. Esta medida se aplica en ontologías taxonómicas prescindiendo del corpus. Mediante la evaluación con las pruebas de rendimiento de Rubensein & Goodenough y Miller & Charles, los resultados muestran que esta medida tiene una precisión prometedora ya que se tiene una mayor correlación a comparación con otras medidas propuestas en la literatura que también son basadas en ontologías.

2.1.3. Medidas de similitud semántica basadas en contenido de información

En [20] se introduce una medida de similitud semántica que combina la propuesta del camino más corto entre un concepto a otro con el contenido de información, que es la probabilidad de ocurrencia de un concepto en un corpus. Esta combinación hace que se mejore la propuesta tradicional del camino más corto añadiendo el contenido de información desde el corpus a su factor de decisión. Esta propuesta supera a los métodos de similitud como lo propuesto por Resnik y en general a las medidas basadas en estructura, dando un valor mayor de correlación.

En [23] se propone una medida de similitud en base al contenido de la información que se puede aplicar a dominios diferentes. Esta medida pretende lograr dos objetivos: el primer objetivo es la universalidad, que se refiere a la posibilidad de aplicarse a muchos dominios diferentes, siempre y cuando el dominio tenga un modelo probabilístico. El segundo objetivo es la justificación teórica, que quiere decir, que la medida sea definida con un conjunto de suposiciones sobre la similitud y no con una sola fórmula.

En [37] se plantea una medida de similitud semántica para una taxonomía con relaciones "is-a" que se basa en el contenido de información compartida. La evaluación experimental indica que esta medida tiene mejores resultados comparados al enfoque de conteo de aristas. Se plantea que en un conjunto de conceptos de una taxonomía de tipo "is-a", la clave para obtener la similitud entre dos conceptos es la información que comparten entre ellos, con este enfoque se dice que mientras más abstracto es un concepto, menor es el contenido de información, entonces mientras más contenido de información compartan, más similares son estos conceptos.

En [27] se revisaron los enfoques existentes de similitud semántica basados en contenido de información desarrollados para los campos de biomédicina y bioinformática, y se propuso un framework unificado para las medidas de similitud semántica en Gene Ontology. Además, se proporcionan los fundamentos teóricos para comparar los diferentes enfoques. Dentro de este framework, se propone una forma de calcular la similitud semántica en base al contenido de información, llamado Nunivers. Los resultados experimentales muestran que la medida propuesta tiene un buen rendimiento a comparación de modelos normalizados basados en la propuesta de Resnik.

En [63] se presenta una nuevo método para calcular el contenido de información que se basa solamente en la estructura jerárquica. Este método considera la profundidad de la estructura y las hiponimias de cada sentido de la palabra. Su principal ventaja es que no se basa en el análisis del corpus, evitando el problema de escasez de datos o a la auscencia de corpus. Este método se puede aplicar a taxonomías como WordNet, Gene Ontology y otras bases de conocimiento.

En [64] se define un método de similitud semántica en grafos de conocimiento, llamado *wpath*, en este trabajo se menciona que hay métodos que miden la similitud semántica en base a la distancia del camino más corto entre dos conceptos de una taxonomía. Por otra parte, hay métodos que toman en consideración el contenido de información (CI) de los conceptos desde el corpus para mejorar el resultado de similitud semántica. A raíz de esto, se propone combinar la propuesta que mide la distancia entre conceptos y el contenido de información, éste se calcula desde el grafo de conocimiento y no desde un corpus. Mediante los resultados experimentales se obtiene que el CI basado en gráfos tiene un rendimiento similar al CI basado en corpus, además, el CI basado en gráfos tiene la ventaja de no requerir el corpus. Al realizar una evaluación, se obtuvo que wpath tiene un mejor rendimiento en precisión en comparación a métodos propuestos por Wu & Palmer, Li, Resnik, Lin y Jiang & Conrath.

2.1.4. Evaluación de ontologías

En [61] se propone un nuevo framework de medición taxonómica en ontologías basado en la relación semántica léxica, se emplean técnicas de conteo de aristas y se hace uso de WordNet para calcular la relación semántica entre dos conceptos en la taxonomía de la ontología. En el framework se utilizaron los algoritmos de búsqueda en profundidad y búsqueda a lo ancho para obtener los pares de conceptos en la ontología y calcular su relación semántica. Se tomaron en cuenta tres tipos de relaciones “is-a”, “part-of” y sinónimo/antónimo.

En [8] se presenta una implementación basada en las reglas de manejo de restricciones con disyunción (CHR, Constraint Handling Rules) que es un lenguaje general de lógica concurrente para la evaluación de ontologías. Esta implementación es una versión mejorada de OntoClean, una metodología para el análisis de ontologías. Esta

propuesta es útil para evaluar ontologías que están parcialmente etiquetadas dado que CHR permite encontrar las propiedades de otros conceptos.

En [54] se propone un framework abstracto para la evaluación de ontologías. Se realizó un análisis de los enfoques propuestos en la literatura para conocer las características de la evaluación de ontologías que se emplean tales como las dimensiones, clasificaciones y métricas. Haciendo este análisis, se propone un framework generalizado usando dimensiones como verificación, validación, exactitud y calidad de la ontología, evaluación basada en usuario/experto, calidad semántica, calidad sintáctica entre otros. Además, algunas de las categorías de las métricas utilizadas son para medir: el tamaño, las clases, las relaciones, la jerarquía, el dominio, y los atributos.

Por otro lado, en Tovar et al. [7, 48–51, 56–58] se ha llevado a cabo la evaluación o validación de relaciones semanticas en ontologías, por medio de enfoques basados en patrones o por análisis formal de conceptos utilizando corpora de dominio.

Tabla 2.1: Tabla comparativa de las medidas de similitud semánticas

Autor(es)	Medida	Categoría	Aspectos que considera
Rada et al. [36]	Path	Basadas en estructura	Camino más corto
Wu y Palmer [62]	WUP	Basadas en estructura	Camino más corto, LCA
Li et al. [21]	Li	Basadas en estructura	Camino más corto, LCA
Maedche y Staab [25]	CMatch	Basadas en características	Unión e intersección de conjuntos
Rodriguez y Egenhofer [38]	RE	Basadas en características	Intersección y diferencia de conjuntos
Sánchez et al. [47]	San	Basadas en características	Intersección y diferencia de conjuntos

Continúa en la siguiente página

Autor(es)	Medida	Categoría	Aspectos que considera
Resnik [37]	Res	Basadas en contenido de información	Contenido de información, MICA
Lin [23]	Lin	Basadas en contenido de información	Contenido de información, MICA
Jiang y Conrath [20]	JC	Basadas en contenido de información	Contenido de información, MICA
Mazandu y Mulder [27]	Nunivers	Basadas en contenido de información	Contenido de información, MICA, valor máximo
Zhou et al. [63]	Zhou	Híbridas	Camino más corto, profundidad máxima, IC, MICA
Paul et al. [32]	HSS3	Híbridas	Camino más corto, IC, MICA
Paul et al. [32]	HSS3	Híbridas	Camino más corto, IC, MICA

Capítulo 3

Marco teórico

En este capítulo, se presentan los fundamentos teóricos necesarios y recomendados para que se pueda comprender mejor el trabajo realizado en esta investigación.

Primero, se iniciará con la definición de Procesamiento de Lenguaje natural, despúes se detalla el concepto de ontología. Posteriormente, se explica la semántica seguido de la similitud semántica y las diferentes medidas que se usarán en este trabajo. Por último, se explican las herramientas que fueron usadas en el sistema propuesto.

3.1. Procesamiento de Lenguaje Natural

El lenguaje natural es la manera en la que los seres humanos expresamos nuestros sentimientos o pensamientos en forma de lenguaje, todo lo que hablamos, escribimos, leemos y escuchamos lo hacemos en lenguaje natural. El procesamiento de lenguaje natural (PLN), según [22] se define como:

"El procesamiento de lenguaje natural es una gama de técnicas computacionales para analizar y representar textos que se producen de forma natural en uno o más niveles de análisis lingüístico con el fin de lograr un procesamiento del lenguaje similar al humano para una variedad de tareas o aplicaciones."

Se dice que el PLN es una gama de técnicas computacionales dado que hay varios métodos que se pueden emplear para llevar a cabo el análisis del lenguaje. Los textos a analizar pueden ser en cualquier lenguaje humano, en formato oral o escrito. El

PLN utiliza diferentes niveles para lograr analizar el lenguaje dependiendo de la aplicación del sistema donde se utilice. El PLN va de la mano con la inteligencia artificial, esto es necesario para hacer que el procesamiento del lenguaje sea parecido al lenguaje humano.

Según [22] los origenes del PLN vienen de diferentes disciplinas:

- Lingüística: se enfoca en modelos formales y estructurales del lenguaje y el descubrimiento de universales lingüísticos.
- Ciencias de la computación: enfocada al desarrollo de representaciones internas de datos y procesamiento eficiente de esas estructuras.
- Psicología cognitiva: considera el uso del lenguaje como ventana hacia el proceso cognitivo del ser humano en la adquisición del conocimiento.

3.1.1. Niveles del procesamiento de lenguaje natural

En [22] los niveles del PLN se presentan de manera secuencial, y es una forma análoga de cómo los seres humanos usan todos estos niveles para la compresión del lenguaje. A continuación se presentan los niveles:

- Fonología: Este nivel se enfoca en la interpretación de sonidos del habla. En un sistema PLN se analizan las ondas de sonido y se codifican en una señal digital para relacionar los sonidos con las palabras que representan.
- Morfología: Este nivel se enfoca en la composición natural de las palabras, que se componen por morfemas (unidad mínima de significado). Se analiza cómo las palabras se integran a partir de los morfemas.
- Léxico: Este nivel analiza el léxico de las palabras, es decir, el conjunto de información sobre cada palabra.
- Sintáctico: Este nivel se enfoca en analizar las palabras para formar una oración.
- Semántico: En este nivel se analiza el significado de las palabras por sí solas y también del significado de las palabras para dar un significado a una oración. Dentro de la semántica, existen diferentes tipos de relaciones semánticas, concepto importante en la investigación de esta tesis (ver sección 3.4.1).

- Discurso. Este nivel analiza con unidades de texto mayores a una oración. Dos tipos de procesamiento de discurso ocurren en este nivel: resolución de anáforas: donde se reemplaza por ejemplo los pronombres por la entidad a la que se refiere; reconocimiento de la estructura del discurso: determina la función de las oraciones en el texto.

- Pragmático: Este nivel analiza la forma en la que las oraciones son usadas en distintas situaciones y cómo afecta su uso en el significado de las oraciones. Usa el contexto por encima del contenido del texto para su comprensión.

3.1.2. Aplicaciones del procesamiento de lenguaje natural

Las aplicaciones más usadas del procesamiento de lenguaje natural son las siguientes:

- Recuperación de información: se encarga de procesar documentos con texto no estructurado para facilitar la búsqueda en el texto a base de palabras clave y retornar al usuario una respuesta entendible.

- Extracción de información: se encarga de extraer información de una manera estructurada a partir de documentos no estructurados.

- Búsqueda de respuestas: Provee al usuario respuestas a preguntas realizadas en lenguaje natural.

- Resumen: se enfoca en los niveles altos del PLN, particularmente en el nivel de discurso, donde desde un texto grande se puede convertir a un texto pequeño, formando un resumen del texto grande.

- Traducción automática: básicamente, como su nombre lo indica, se enfoca en traducir texto o voz de un lenguaje a otro.

- Sistemas de diálogo: tienen como objetivo facilitar la interacción entre un humano y una máquina. La máquina obtiene la voz emitida por un humano, la procesa y proporciona respuestas en lenguaje natural.

3.2. Ontologías

En [15] se define a una ontología como "una especificación explícita de una conceptualización". Posteriormente, [6] definió una ontología como "una especificación formal de una conceptualización compartida".

Teniendo estas dos definiciones, [46] definió e hizo una explicación de la misma, "una ontología es una especificación formal y explícita de una conceptualización compartida".

Conceptualización se refiere a un modelo abstracto de un fenómeno en el mundo al haber identificado los conceptos relevantes de ese fenómeno. Explícito significa que el tipo de conceptos y restricciones usados están explícitamente definidos. Formal se refiere al hecho en que la ontología debe ser legible para una máquina. Compartida refleja la noción que una ontología captura el conocimiento consensuado, esto es, que no es privado de un individuo, sino aceptado por un grupo".

Las ontologías son modelos formales que describen un dominio en particular y especifican los conceptos dentro de ese dominio con relaciones entre ellos, las ontologías son dinámicas, es decir, cambian conforme pasa el tiempo.

3.2.1. Principales componentes de una ontología

Gruber [15] identificó cinco tipos de componentes que se detallan a continuación:

- Clases: representan conceptos, por ejemplo, en el dominio de transporte, los conceptos serían carros, trenes, aviones, etc. Las clases en una ontología usualmente están organizadas en taxonomías en donde se aplican mecanismos de herencia. Las clases pueden representar conceptos abstractos o conceptos específicos. En la Figura 3.1, son representadas con el círculo amarillo: Persona, Ingeniero y Licenciado.

- Relaciones: representan un tipo de asociación entre los conceptos del dominio. Las ontologías usualmente usan relaciones binarias, en donde el primer elemento es el dominio de la relación y el segundo es el rango. En la Figura 3.1, las relaciones son las representadas con la flecha color azul que representa una relación de inclusión de tipo hiperonimia/hiponimia (ver sección 3.4.1).

- Funciones: son un tipo de relación donde se identifica un elemento haciendo un cálculo de una función donde se considera diferentes elementos de la ontología.
- Axiomas formales: sirven para modelar oraciones que son siempre verdaderas, son usadas para representar conocimiento que no puede ser representado formalmente por otros componentes. También son usadas para verificar la consistencia de la ontología. Son útiles para inferir nuevo conocimiento.
- Instancias: Son usadas para representar elementos o individuales en una ontología. En la Figura 3.1, son las representadas con el rombo púrpura: Ing_Israel_Hernandez y Lic_Ernesto_Lopez.

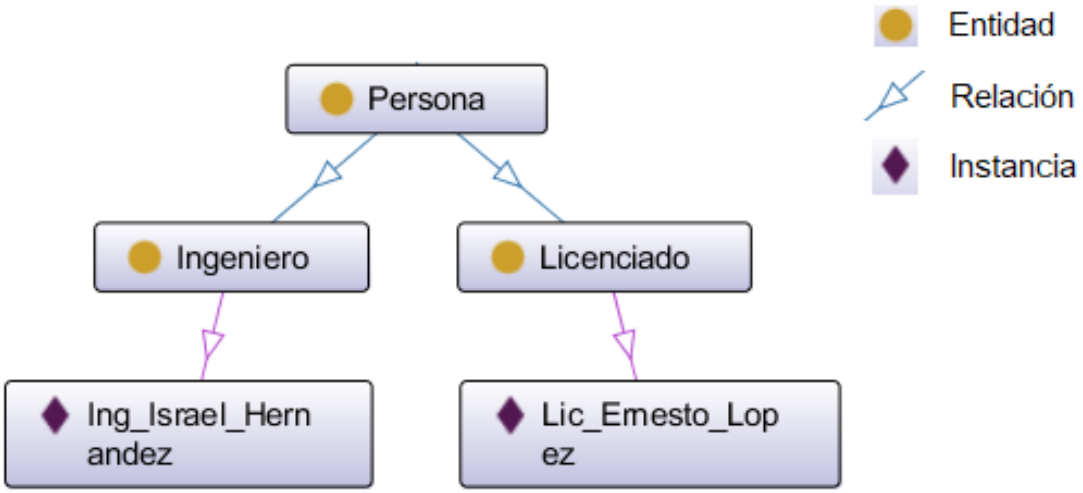

Figura 3.1: Componentes en una ontología.

3.2.2. Clasificación de ontologías

Una clasificación de las ontologías se introdujo en [17] donde se usa la conceptualización como criterio principal y se sugiere desarrollar diferentes tipos de ontologías de acuerdo a su nivel de generalidad, como se ve en la Figura 3.2.

En donde cada tipo de ontología se define como:

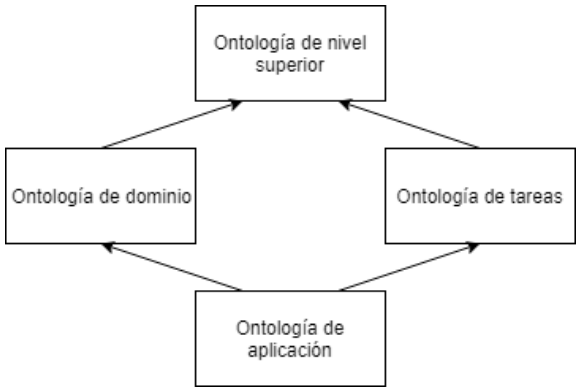

Figura 3.2: Tipos de ontologías y sus relaciones [17].

- Ontologías de nivel superior: describen conceptos muy generales que pueden ser aplicadas a varios dominios. También se conocen como ontologías fundacionales.

- Ontologías de dominio: describen el vocabulario relacionado a un dominio genérico mediante la especialización de los conceptos introducidos en la ontología de nivel superior.

- Ontologías de tareas: describen el vocabulario relacionado a una tarea o actividad genérica mediante la especialización de las ontologías de nivel superior.

- Ontologías de aplicación: son las ontologías más específicas. Los conceptos en estas ontologías a menudo corresponden a roles desempeñados por entidades de dominio mientras se realiza una actividad determinada.

3.3. Ingeniería ontológica

Es una disciplina que investiga los métodos y herramientas para el desarrollo y mantenimiento de las ontologías. El objetivo de esta disciplina es proporcionar indicaciones y consejos a los desarrolladores de ontologías. Toda la información sobre esta sección ha sido recopilada del trabajo de [45].

Aprendizaje automático de ontologías

Todo lo que implica el diseño de una ontología es una tarea costosa que consume muchos recursos. Para reducir los costos se necesitan de técnicas que ayuden a la tarea de la ingeniería ontológica. El uso incremental de las ontologías en áreas como la web semántica, la inteligencia artificial, biomedicina entre otras, hace que en la literatura se encuentren cada vez más propuestas de investigadores con métodos que facilitan la tarea de diseño de ontologías. El conjunto de estos métodos es conocido como aprendizaje automático de ontologías.

El aprendizaje automático de ontologías se basa en técnicas con fundamentos en diferentes disciplinas como las son: el procesamiento de lenguaje natural, el aprendizaje automático, la adquisición de conocimiento y la ingeniería ontológica. El proceso general del aprendizaje de ontologías se desarrolla de manera semiautomática con intervención humana dado que actualmente no existe la adquisición completamente automática del conocimiento.

Arquitectura y modelo

En este apartado se presenta una arquitectura genérica para el aprendizaje automático de ontologías y también sus principales componentes que la integran.

Esta arquitectura cuenta con cuatro componentes principales: una gestión de ontologías, una coordinación, un procesamiento de recursos y una biblioteca de algoritmos que se describen a continuación:

- Componente de gestión de ontología: este componente tiene como objetivo la manipulación de la ontología, facilitando así las tareas de importación, navegación, modificación y control de versiones. Ese componente proporciona una interfaz al ingeniero ontológico entre la ontología y los algoritmos de aprendizaje.

- Componente de coordinación: el objetivo de este componente es organizar y aplicar secuencialmente los algoritmos que fueron seleccionados, así como los métodos de procesamiento de recursos.

- Componente de procesamiento de recursos: en este componente se genera el conjunto de datos preprocesados que se pasarán al componente de biblioteca

de algoritmos. Este componente contiene una gama de técnicas para analizar y transformar los datos de entrada, teniendo como subcomponente principal el procesamiento de lenguaje natural.

- Componente de biblioteca de algoritmos: este componente es principal porque aquí se ofrecen varios algoritmos dedicados a la extracción y mantenimiento de las primitivas del modelado de ontologías. Esta biblioteca consta de algoritmos de aprendizaje automático y versiones de éstas que son aplicables al aprendizaje de ontologías.

Evaluación de ontologías

La evaluación de ontologías es un procedimiento esencial en el aprendizaje de ontologías ya que busca evaluar la calidad de las ontologías desarrolladas. Este procedimiento puede ayudar a orientar y controlar el proceso del aprendizaje de ontologías porque se especializa en la búsqueda de una ontología óptima.

Clasificación de evaluación de ontologías

La clasificación de las propuestas en la evaluación de ontologías se basan en los criterios que se consideran y en los motivos por los cuales fueron diseñadas. De acuerdo a Raad y Cruz [35], estas propuestas se clasifican de la siguiente manera:

- Basadas en estándar de oro: en este tipo de evaluación, se compara la ontología creada con un estándar de oro que puede ser otra ontología creada por un experto como en el trabajo de [25].

- Basadas en corpus: en esta evaluación se compara la ontología con un corpus textual que cubre significativamente el dominio dado para determinar hasta que punto la ontología cubre ese dominio.

- Basadas en tareas: este tipo de evaluación se aplica cuando la ontología fue diseñada para una tarea en específico donde se evalúa su desempeño al ejecutar dicha tarea.

- Basadas en criterios: como su nombre lo indica, este tipo de evaluación se realiza teniendo criterios los cuales la ontología debe cumplir, estos criterios se basan en la estructura de la ontología.

A continuación se discutirán los criterios de calidad ontológica según la literatura.

Criterios

En la evaluación de ontologías existen diferentes tipos de criterios que describen los principios de una buena ontología. Usualmente no se eligen todos los criterios que se mencionan en la literatura, estos criterios son elegidos por el ingeniero ontológico considerando aquellos que son adecuados con la finalidad de la ontología.

Gómez-Pérez[14] presenta dos términos para la evaluación de ontologías: la verificación y la validación. La verificación se asegura de que las definiciones cumplan los requisitos correctamente. Por otra parte, la validación se asegura que el significado de las definiciones modelen correctamente los aspectos del mundo.

Aparte de la verificación y validación, se discute también los siguientes criterios de calidad en la literatura.

- Exactitud: en [31] la exactitud es cuando la ontología puede capturar y representar los aspectos del mundo de una manera correcta.

- Adaptabilidad: en [31] la adaptabilidad se refiera a si la ontología ofrece una base conceptual para tareas anticipadas. En [14] adaptabilidad es observar cómo la ontología reacciona a cambios en los axiomas.

- Claridad: el criterio de claridad en [15] se asegura que la ontología comunique de manera eficaz el significado de los términos definidos, en general, si la ontología es comprensible.

- Integridad/Competencia: en la integridad y/o competencia según [14] y [16] se verifica si el dominio de la ontología está cubierto adecuadamente. También comprueba si se han definido las preguntas de competencia y si la ontología es capaz de responderlas. Se verifica que la ontología incluya todos los conceptos relevantes del dominio junto con sus representaciones léxicas.

- Eficiencia computacional: En [9] y [31] la eficiencia computacional se refiere a la facilidad que tienen los razonadores para procesar la ontología, así como a la rapidez en la que se pueden aplicar los servicios de razonamiento.

- Concisión: En [14] en la concisión se examina si la ontología incluye axiomas redundantes o irrelevantes con respecto al dominio.

- Consistencia/Coherencia: En [14] y [15] este criterio es definido para saber si la ontología posee consistencia lógica, es decir, si no existen axiomas que conduzcan a la contradicción.

- Accesibilidad comercial: En [9] este criterio se define para saber si en una ontología tiene una certificación y si es de fácil acceso. Otro aspecto importante es si la ontología puede alinearse con otras que están en uso.

Aspectos

A continación, se definen diferentes aspectos que son susceptibles a la evaluación de una ontología.

- Vocabulario: se refiere al conjunto de todos los nombres en la ontología, pueden ser URI o literales.

- Sintaxis: la sintaxis de una ontología puede estar escrita en alguno de los siguientes formatos: RDF/XML, N-triples, OWL, etc.

- Estructura: la estructura de una ontología está representada en un grafo.

- Semántica: la semántica se refiere a las características comunes del conjunto de modelos.

- Representación: se refiere a la relación entre la estructura y la semántica.

- Contexto: Este aspecto trata sobre las características de la ontología cuando se compara con otros artefactos en su entorno, como puede ser contra las preguntas de competencia.

3.4. Semántica

En [55] se define a la semántica como la disciplina que estudia el significado de las expresiones lingüísticas. Dentro de la semántica, se estudian dos tipos: la semántica léxica y la semántica composicional.

- Semántica léxica: se ocupa de todo lo relacionado al significado de las palabras.
- Semántica composicional: se centra en descubrir cuál es la contribución que la estructura y las relaciones sintácticas aportan a la construcción de las expresiones complejas.

3.4.1. Relaciones semánticas

En esta sección se centrará en explicar solamente las relaciones de inclusión que son las relaciones con las que se trabajará en esta investigación, especialmente con las relaciones hiponimia/hiperonimia. En [55] se menciona que una de las propiedades notables del léxico de una lengua es que los significados de las palabras están conectados por diferentes tipos de relaciones.

Identidad y semejanza

Se define sinonimia a la relación que se establece entre dos palabras que tienen el mismo significado, por ejemplo, computadora y ordenador significan lo mismo, por consiguiente, son palabras sinónimas. Entonces la sinonimia es una relación de identidad entre significados. Dos características de la sinonimia es que los dos significados tienen el mismo contenido semántico y que las los palabras sinónimas pueden intercambiarse sin que se modifique el significad del texto.

Inclusión

Se dice que hay inclusión cuando el significado de una palabra está contenido en el significado de otra. Un ejemplo es, rosa es un tipo de flor, entonces se puede decir que rosa está contenido en el significado de flor. Este tipo de relaciones se clasifican en las siguientes categorías:

- Hiponimia/hiperonimia: esta relación se presenta entre un hipónimo y un hiperónimo, en el ejemplo anterior, el hipónimo es rosa y el hiperónimo es flor, lo que quiere decir que un hiperónimo posee un significado más general que el hipónimo, que posee un significado más específico. Entonces se puede decir que el hiperónimo incluye al hipónimo. Como ejemplo, observando la Figura 3.3, se tiene que además de una rosa, también existen diferentes tipos de flores, como tulipanes, orquídeas, lirios, etc. entre estos términos también existe una relación que se llama co-hiponimia y cada término se denominan como co-hipónimos. La relación de inclusión es de tipo transitiva, por ejemplo, que el significado de flor incluye al significado de rosa, entonces el significado de planta incluye al de flor, por consiguiente, el significado de planta incluye al significado de rosa. La relación hiponimia/hiperonima se puede apreciar gráficamente en la Figura 3.3

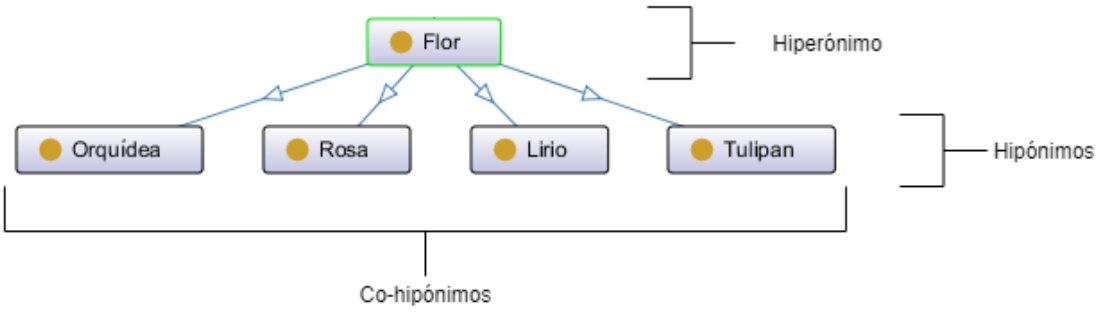

Figura 3.3: Representación de la relación entre hiperónimo-hipónimo.

- Meronimia/holonimia: este tipo de relación de basa en la relación parte-todo. Un ejemplo de esto, es un *mouse* es parte de una computadora de escritorio, entonces existe una relación de inclusión entre *mouse* y computadora de escritorio. Por lo tanto, meronimia es la relación que existe entre los significados de estos términos. Se le llama merónimo al término incluido, en el ejemplo anterior, el merónimo es *mouse*, el holónimo es el término incluyente, como lo es computadora de escritorio.

Exclusión

La oposición tiene lugar cuando la relación de exclusión tiene las siguientes características: binarias, inherentes y patentes.

Los tipos de relaciones de oposición son las siguientes:

- Complementarios: aquí se clasifican los términos con significados mutuamente excluyentes donde no existe un término medio, es decir, no son graduables. Ejemplos: verdadero-falso, vivo-muerto, etc.

- Antónimos: son términos en los que si existe una escala graduable. Se dividen en dos subtipos:

  - Antónimos polares: sus propiedades se pueden medir de manera objetiva, además admiten modificadores como lo son: poco, mucho, bastante, etc. Ejemplos: grande - pequeño, fuerte - débil, etc.
  - Antónimos equipolentes: estos términos describen sensaciones o emociones. Ejemplos: frío-caliente, dulce-salado, triste-alegre, etc.

- Inversos: estos términos expresan una relación desde perspectivas opuestas que tienen la característica de describir la misma situación sin dejar de ser opuestos. Ejemplos: encima-debajo, comprar-vender, etc.

- Reversos: este tipo de términos indican movimiento en donde existe un cambio de estado en dirección opuesta. Ejemplos: entrar-salir, abrir-cerrar, atar-desatar, etc.

Jerarquías taxonómicas

Las jerarquías coinciden con las relaciones semánticas de inclusión. Por lo tanto, las jerarquías taxonómicas se crean a partir de la relación de hiponimia (A es un tipo de B) por lo que las taxonomías o jerarquías taxonómicas son redes de relaciones de hiponimia. En la Figura 3.4 se puede apreciar un ejemplo sencillo de una taxonomía.

En una taxonomía existen términos más básicos que otros, estos términos se suelen posicionar en los niveles más bajos. Estos términos básicos tienen características propias que pueden diferenciarlos de términos de otros niveles. Como se parte de

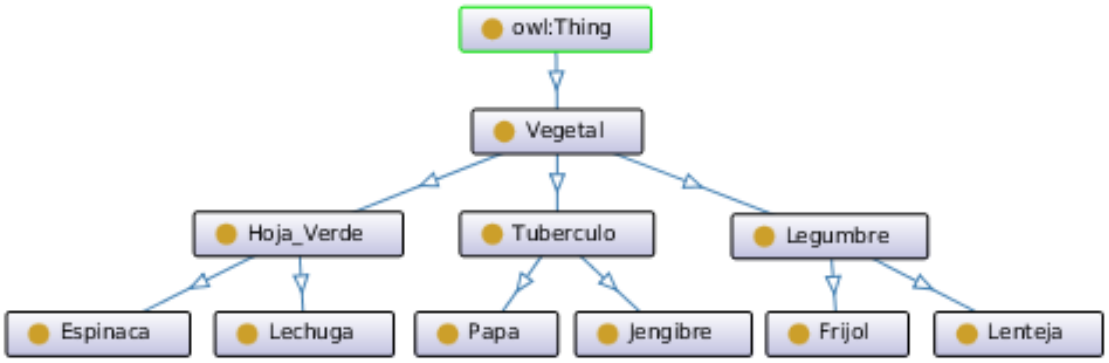

Figura 3.4: Ejemplo de una taxonomía

las relaciones de hiponimia, los términos inferiores en la taxonomía están formados por expresiones más complejas que a los términos superiores, que tienden a tener expresiones simples. Una de las características principales de una taxonomía y una forma para comprobar su validez, es que tienen un principio de unicidad de dominio, lo que quiere decir, que una categoría está dominada por sólo una categoría superior, no debe estar dominada por dos o más categorías superiores.

3.5. Similitud semántica

En el trabajo de Pedersen et al. [33] se define que "La similitud es un caso especial de relación semántica que está ligada a la semejanza de los conceptos. Una medida de similitud semántica toma como entrada dos conceptos y devuelve una puntuación numérica que cuantifica en qué medida se parecen. Dicha medida generalmente se basa en las relaciones is-a que se encuentran en la taxonomía u ontología subyacente en la que residen los conceptos". Harispe et al. [18] define la similitud semántica como "subconjunto de la noción de relación semántica que sólo tiene en cuenta las relaciones taxonómicas en la evaluación de la interacción semántica entre dos elementos".

3.5.1. Medidas de similitud semántica

Harispe et al. [18] define las medidas de similitud semántica como "herramientas matemáticas utilizadas para estimar la fuerza de la relación semántica entre unidades de lenguaje, conceptos o instancias, a través de una descripción (numérica) obtenida de acuerdo con la comparación de información que respalda su significado".

3.5.2. Clasificación de las medidas de similitud semántica

En [18] se clasifican las medidas en dos categorías:

- Medidas basadas en corpus: son las medidas que comparan unidades del lenguaje, ya sean conceptos o instancias desde el análisis de un texto sin estructurar.
- Medidas basadas en conocimiento: estas medidas están diseñadas para comparar entidades definidas en una ontología o recursos estructurados.

En esta investigación nos centramos únicamente en las medidas de similitud semántica basadas en conocimiento ya que son las medidas que pueden ser utilizadas en taxonomías y/o ontologías.

3.5.3. Medidas basadas en conocimiento

Este tipo de medidas son usadas generalmente para comparar términos estructurados en relaciones semánticas conceptos definidos en taxonomías. Estas medidas evalúan la similitud semántica de un par de conceptos definidos en la taxonomía. Los principales enfoques propuestos en la literatura de este tipo de medidas se presentan a continuación.

Enfoque basado en estructura

Estas medidas estiman la similitud de acuerdo a la distancia de dos conceptos en la ontología, es decir, se calcula la longitud de los caminos que conectan a los conceptos que son comparados.

A continuación se describen las medidas de similitud que se implementaron para esta investigación.

Rada En [36] se propone una medida llamada Distancia, y en base a esta medida en el trabajo de [64] se definió la ecuación sim_{Path} (ver Ecuación 3.1), usando el camino más corto (*length*) entre dos conceptos c_u y c_v, mediante esta distancia se puede saber la similitud entre los conceptos en una taxonomía.

$$sim_{path}\left(c_u,\ c_v\right) = \frac{1}{1 + length\left(c_u,\ c_v\right)} \tag{3.1}$$

Como ejemplo se toma la taxonomía de la Figura 3.4 los conceptos son c_u =*Frijol* y c_v =*Legumbre*, como se puede observar el camino más corto o *length* entre estos conceptos es de 1, entonces aplicando la Ecuación 3.1 para este par de conceptos queda de la siguiente manera.

$$sim_{path}\left(Frijol,\ Legumbre\right) = \frac{1}{1+1} = 0.5 \tag{3.2}$$

Para las siguientes medidas presentadas en el trabajo de [64] se necesitan dos conceptos para implementarlas, primeramente se define el concepto de profundidad o *depth*. La profundidad de un concepto ($depth(c_u)$) es el camino más corto desde un concepto c_u al concepto raíz $c_{raíz}$ y se define en la Ecuación 3.3.

$$depth\left(c_u\right) = length\left(c_u,\ c_{raíz}\right) \tag{3.3}$$

El segundo concepto es el subsumidor menos común o *Least Common Subsumer (LCS)* también conocido como el ancestro menos común *Least Common Ancestor (LCA)* que en [64] se define como el concepto más específico que es ancestro común de dos conceptos.

Wu y Palmer En base a las definiciones antes presentadas y a la investigación de [62] se define la medida de Wu y Palmer en [64], donde se mide la similitud de dos conceptos c_u y c_u donde se mide la distancia más corta de cada concepto c_u y c_v con el concepto raíz y la distancia del LCA de cada concepto c_u y c_v con el concepto raíz. Esta medida se define en la Ecuación 3.4.

$$sim_{wup}\left(c_u,\ c_v\right) = \frac{2 * depth\left(c_{LCA}\right)}{depth(c_u) + depth(c_v)} \tag{3.4}$$

En donde $depth(c_i)$ y $depth(c_j)$ es la distancia más corta de cada concepto con el concepto raíz y $depth(c_{LCA})$ es la distancia del LCA con el concepto raíz.

Como ejemplo, para los conceptos c_i =*Frijol* y c_j =*Legumbre* (ver Fig. 3.4) donde el LCA es *Legumbre*, el valor de $depth(c_{LCA})$ es de 3, mientras que los valores para $depth(Frijol) = 4$ y $depth(Legumbre) = 3$, si $depth(c_{raíz}) = 1$, entonces el cálculo de la similitud semántica utilizando la Ecuación (3.4) es de 0.857, ver resultado en la Ecuación (3.5).

$$sim_{wup}\,(Frijol,\ Legumbre) = \frac{2 * 3}{4 + 3} = 0.857 \tag{3.5}$$

Li En base al método diseñado por Li [21] se formula la Ecuación 3.6 propuesta por [64], donde se combina el camino más corto(*depth*) de ambos conceptos c_u y c_v y el LCA (c_{LCA}) de los conceptos, para calcular su similitud.

$$sim_{li}\,(c_u,\ c_v) = e^{-\alpha length(c_u,c_v)} \cdot \frac{e^{\beta depth(c_{LCA})} - e^{-\beta depth(c_{LCA})}}{e^{\beta depth(c_{LCA})} + e^{-\beta depth(c_{LCA})}} \tag{3.6}$$

Donde α es un parámetro que contribuye a la longitud del camino y β es el parámetro para la profundidad del camino. De acuerdo con el trabajo de [21], el parámetro óptimo para α es 0.2 y para β es 0.6.

Por ejemplo, para los conceptos c_u = *Frijol* y c_v = *Legumbre* (ver Fig. 3.4), el valor para $length(c_u, c_v)$ es de 1 y el valor para $depth(c_{LCA})$ es de 3, con estos valores se calcula la similitud semántica para este par de conceptos y se obtiene el resultado de 0.775 (ver Ecuación 3.7).

$$sim_{li}\,(Frijol,\ Legumbre) = e^{-0.2*1} \cdot \frac{e^{0.6*3} - e^{-0.6*3}}{e^{0.6*3} + e^{-0.6*3}} = 0.3745 \tag{3.7}$$

Enfoque basado en características

Estas medidas primero extraen características de los conceptos en la ontología. Para estimar la similitud semántica, se analizan las características que comparten o que son diferentes entre los conceptos que están siendo comparados.

Maedche Esta medida propuesta en [25] a partir del índice de Jaccard, considera las características o el conjunto de ancestros de un concepto que se define como

$$sim_{CMatch}(u,v) = \frac{|A(u) \cap A(v)|}{|A(u) \cup A(v)|} \tag{3.8}$$

La Ecuación 3.8 fue propuesta en el trabajo [18].

Como ejemplo, se toman de nuevo los dos conceptos $u = Frijol$ y $v = Legumbre$ (ver Fig. 3.4). Para $Frijol$ el conjunto de ancestros es $A(u) = \{Vegetales, Legumbre, raíz\}$ para $Legumbre$ su conjunto $A(v) = \{Vegetales, raíz\}$, la intersección de estos dos conjuntos resulta $A(u) \cap A(v) = \{raíz, Vegetales\}$ siendo 2 elementos y la unión es $A(u) \cup A(v) = \{raíz, Vegetales, Legumbre\}$ con 3 elementos. Entonces utilizando la Ecuación 3.8 el resultado de similitud semántica para la medida propuesta por Maedche se muestra en la Ecuación 3.9

$$sim_{CMatch}(Frijol, Legumbre) = \frac{|A(Frijol) \cap A(Legumbre)|}{|A(Frijol) \cup A(Legumbre)|} = \frac{2}{3} = 0.667 \tag{3.9}$$

Rodriguez y Egenhofer Esta medida fue propuesta por [38] es una normalización del modelo de [53] con funciones de teoría de conjuntos como la intersección ($\cap$) y la diferencia ($\setminus$) . La Ecuación 3.10 es la propuesta por [38] siguiendo el formato de [18].

$$sim_{RE}(u,v) = \frac{|A(u) \cap A(v)|}{\gamma|A(u) \setminus A(v)| + (1-\gamma)|A(v) \setminus A(u)| + |A(u) \cap A(v)|} \tag{3.10}$$

Donde γ es un parámetro para ajustar la simetría de esta medida donde $\gamma \in [0,1]$, $A(u)$ es el conjunto de ancestros del concepto u y $A(v)$ es el conjunto de ancestros del concepto v.

Se toman los dos conceptos $u = Frijol$ y $v = Legumbre$ y el concepto *raíz*= *owl:Thing* (ver Fig. 3.4). Para $Frijol$ el conjunto de ancestros es $A(u) = \{Vegetales, Legumbre, raíz\}$ para $Legumbre$ su conjunto $A(v) = \{Vegetales, raíz\}$, la intersección de estos dos conjuntos resulta $A(u) \cap A(v) = \{raíz, Vegetales\}$ siendo 2 elementos. La diferencia $A(u) \setminus A(v) = \{Legumbre\}$ y $A(v) \setminus A(u) = \{\}$, es decir, un conjunto vacío. Para este ejemplo $\gamma = 0.2$. Entonces utilizando la Ecuación 3.10 el resultado de similitud semántica para la medida propuesta por Rodriguez y Egenhofer se muestra en la Ecuación 3.11

$$sim_{RE}(Frijol, Legumbre) = \frac{2}{0.2(1) + (1-0.2)(0) + 2} = 0.909 \tag{3.11}$$

Sánchez Propuesta en [47], en esta medida se calcula la distancia taxonómica de dos conceptos y dividiendo la cardinalidad del conjunto de características diferentes con y viceversa, entre la suma de las cardinalidades de los conjuntos de características taxonómicas diferentes y comunes. Esta medida mostrada en la Ecuación 3.12 también sigue el formato de [18].

$$dist_{Sanchez}(u,v) = log_2\left(1 + \frac{|A(u) \setminus A(v)| + |A(v) \setminus A(u)|}{|A(u) \setminus A(v)| + |A(v) \setminus A(u)| + |A(u) \cap A(v)|}\right) \tag{3.12}$$

Para evitar valores infinitos en términos equivalentes al logaritmo se le agrega 1 a la ecuación. Además, esta ecuación tiene un rango de valores en los resultados en un intervalo de $[0, 1]$.

Para esta medida, con los conceptos usados para mostrar un ejemplo son *Frijol* y *Legumbre* (ver Fig. 3.4). Para $Frijol$ el conjunto de ancestros es $A(u) = \{Vegetales, Legumbre, raíz\}$ para $Legumbre$ su conjunto $A(v) = \{Vegetales, raíz\}$, la intersección de estos dos conjuntos resulta $A(u) \cap A(v) = \{raíz, Vegetales\}$ teniendo 2 elementos. La diferencia $A(u) \setminus A(v) = \{Legumbre\}$ con 1 elemento y $A(v) \setminus A(u) = \{\}$, siendo un conjunto vacío. Entonces utilizando la Ecuación 3.12 el resultado de similitud semántica para la medida propuesta por Sánchez se muestra en la Ecuación 3.13.

$$dist_{Sanchez}(Frijol, Legumbre) = log_2\left(1 + \frac{1+0}{1+0+2}\right) = 0.415 \tag{3.13}$$

Enfoque basado en contenido de información

Estas medidas estiman la similitud semántica dependiendo de la cantidad de información que comparten o difieren de los conceptos que están siendo comparados. Las medidas usan el contenido de información intrínsico, es decir, que sólo toma en cuenta la estructura de la taxonomía para su cálculo. La fórmula que se define en 3.14 fue propuesta por Zhou et al. [63] y se basa en la profundidad de la estructura y en el número de descendientes del concepto.

$$IC_c = k\left(1 - \frac{log(|D(c)|)}{log(|C|)}\right) + (1-k)\frac{log(depth(c))}{log(depth(G_T))} \tag{3.14}$$

Donde $|C|$ es el número de conceptos en la ontología, $D(c)$ son los descendientes del concepto c, $depth(c)$ es la profundidad del concepto en la ontología, $depth(G_T)$

es la profundidad máxima de la ontología y $k \in [0, 1]$ es un parámetro para ajustar el peso de la ecuación y se considera $k = 0.6$.

Resnik En [37] se calcula la similitud entre un par de conceptos (u, v) calculando el contenido de información (CI) del ancestro común entre los dos conceptos u y v que tiene el máximo valor de CI entre todos sus ancestros (MICA, Most Informative Common Ancestor) y se define en la Ecuación 3.15.

$$sim_{Resnik}(u, v) = IC(MICA(u, v)) \tag{3.15}$$

Por ejemplo, utilizando los conceptos con los que hemos estado trabajando (ver Fig. 3.4), el contenido de información para este par de conceptos tienen los siguientes valores: $CI(Frijol) = 1$, $CI(Legumbre) = 0.688$, teniendo estos valores, se calcula el CI del concepto MICA, el concepto MICA es el concepto ancestro común del par de conceptos, que para este caso $MICA(Frijol, Legumbre) = Legumbre$, entonces $CI(MICA) = 0.688$. La similitud semántica con la propuesta de Resnik aplicando la equación 3.15 se calcula como se muestra en la Ecuación 3.16.

$$sim_{Resnik}(Frijol, Legumbre) = 0.688 \tag{3.16}$$

Lin En [23] se propone una medida basada en el trabajo de [37] pero agregando el contenido de información del par de conceptos (u, v), esta medida se define en la Ecuación 3.17.

$$sim_{Lin}(u, v) = \frac{2 \cdot IC(MICA(u, v))}{IC(u) + IC(v)} \tag{3.17}$$

Con los datos del ejemplo anterior de los conceptos $Frijol$ y $Legumbre$ (ver Fig. 3.4), $CI(Frijol) = 1$, $CI(Legumbre) = 0.688$, y $CI(MICA) = 0.688$, la similitud semántica para la medida propuesta por Lin, se calcula de la manera como se muestra en la Ecuación 3.18.

$$sim_{Lin}(Frijol, Legumbre) = \frac{2 \cdot 0.688}{1 + 0.688} = 0.815 \tag{3.18}$$

Jiang y Conrath Esta medida de distancia semántica propuesta en [20] se deriva de la propuesta basada en aristas, añadiendo contenido de información a su factor de

decisión. Se toma en cuenta: la fuerza del enlace, es decir, el contenido de información entre el concepto padre y el concepto hijo, el camino más corto entre los dos conceptos (u, v) y el contenido de información del MICA de los mismos. Definiendo la Ecuación 3.19 para medir la distancia semántica.

$$dist_{JC}(u, v) = IC(u) + IC(v) - 2 \cdot IC(MICA(u, v)) \tag{3.19}$$

En el trabajo de [40] la Ecuación 3.19 se normaliza y se aplica transfomación lineal para convertirla en una función de similitud, quedando definida como se muestra en la Ecuación 3.20.

$$sim_{JC}(u, v) = 1 - \left(\frac{IC(u) + IC(v) - 2 \cdot IC(MICA(u, v))}{2} \right) \tag{3.20}$$

Aplicando la Ecuación 3.20 con los conceptos $Frijol$ y $Legumbre$ (ver Figura 3.4), se calcula la similitud semántica como se muestra en la Ecuación 3.21.

$$sim_{JC}(Frijol, Legumbre) = 1 - \left(\frac{1 + 0.688 - 2 \cdot 0.688}{2} \right) = 0.844 \tag{3.21}$$

Mazandu y Mulder Esta medida llamada, Nunivers, propuesta en [27], la similitud semántica entre dos conceptos se obtiene calculando el MICA entre los conceptos (u, v), entre el valor máximo del contenido de información de los conceptos u y v. La medida Nunivers se define en la Ecuación 3.22.

$$sim_{Nunivers}(u, v) = \frac{IC(MICA(u, v))}{max(IC(u), IC(v))} \tag{3.22}$$

Con los datos de contenido de información de los conceptos $Frijol$ y $Legumbre$ (ver Figura 3.4), $CI(Frijol) = 1$, $CI(Legumbre) = 0.688$ se calcula la similitud semántica con la propuesta hecha por Mazandu y Mulder, observar que el valor máximo entre $CI(Frijol) = 1$, $CI(Legumbre) = 0.688$ es $CI(Frijol)$, entonces $max(IC(u), IC(v)) = 1$. Aplicando la Ecuación 3.22 con este par de conceptos se tiene la Ecuación 3.23.

$$sim_{Nunivers}(Frijol, Legumbre) = \frac{0.688}{1} = 0.688 \tag{3.23}$$

Enfoque híbrido

Estas medidas combinan elementos de diferentes enfoques mencionados anteriormente. Los enfoques que más se combinan son los enfoques basados en estructura con los enfoques basados en contenido de información. A continuación, se presentan tres medidas híbridas.

Zhou Esta medida híbrida fue propuesta por [63] para medir la similitud semántica en la taxonomía de WordNet. La medida combina el camino más corto y el contenido de información y se define en la Ecuación 3.24.

$$sim_{zhou}(u,v) = (1-k) * \frac{log(length(u,v)+1)}{log(2*maxDepth(c)-1)} - (1-k) * \frac{dist_{JC}(u,v)}{2} \quad (3.24)$$

Donde k es un factor de peso que puede adaptarse para cambiar el peso de la distancia del camino y el valor de CI. $maxDepth(c)$ es la profundidad máxima de la ontología. En la ecuación 3.25 se presenta el ejemplo con los conceptos *Frijol* y *Legumbre* (ver Figura 3.4). Donde $k = 0.5$, $dist_{JC} = 0.312$, $maxDepth = 3$.

$$sim_{zhou}(Frijol, Legumbre) = (0.5) * \frac{log(2)}{log(5)} - (0.5) * \frac{0.312}{2} = 0.137 \quad (3.25)$$

HSS3 Esta medida propuesta en [32], introduce parámetros de suavizado como la constante K, donde $K = \frac{1}{ND}$ y ND es el número total de desórdenes en la propuesta original, en este caso se tomó el número total de relaciones taxonómicas. Esta medida se define en la Ecuación 3.26.

$$sim_{HSS3}(u,v) = \frac{\frac{L}{D} * (IC(MICA(u,v)) + K)}{length(u, c_{LCA}) + length(v, c_{LCA})} \quad (3.26)$$

En esta ecuación, L es la longitud del camino desde el concepto raíz y el concepto LCA, es decir, $depth(c_{LCA})$ y D es la profundidad de la ontología o $maxDepth(c)$, como en la medida de Zhou 3.24.

Con los conceptos $u =$ *Frijol* y $v =$ *Legumbre* (ver Figura 3.4), en la Ecuación 3.27 se tiene la aplicación de la Ecuación 3.26 y el resultado de su similitud semántica. Donde $K = \frac{1}{9}$, $L = 2$, $D = 3$, $length(u, c_{LCA}) = 1$ y $length(u, c_{LCA}) = 0$, ya que *Legumbre* es el concepto LCA.

$$sim_{HSS3}(Frijol, Legumbre) = \frac{\frac{2}{3} * (0.688 + 0.111)}{1} = 0.532 \tag{3.27}$$

HSS4 Esta medida también fue propuesta en [32], donde es una modificación de la medida HSS3 definida en la Ecuación 3.26 pero con la diferencia de calcular el contenido de información de cada concepto y prescindiendo de la constante K.

$$sim_{HSS4}(u, v) = \frac{\frac{L}{D} * (IC(MICA(u, v)) + \frac{IC(u)*IC(v)}{IC(u)+IC(v)})}{length(u, c_{LCA}) + length(v, c_{LCA})} \tag{3.28}$$

En la Ecuación 3.29, se muestra la aplicación de la medida HSS4 con los conceptos *Frijol* y *Legumbre* (ver Figura 3.4).

$$sim_{HSS4}(Frijol, Legumbre) = \frac{\frac{2}{3} * (0.688) + \frac{1*0.688}{1+0.688})}{1} = 0.730 \tag{3.29}$$

3.6. Herramientas para ontologías

En esta sección se presentan las herramientas principales para la creación y manejo de ontologías, también se presentan las principales bibliotecas para el uso de medidas de similitud semánticas que se utilizaron en esta investigación.

3.6.1. XML (Extensible Markup Language)

XML es un metalenguaje con el que se pretende transformar la web en la web semántica, por esta razón, la mayoría de las tecnologías hechas para la web semántica se basan en este metalenguaje.

XML usa etiquetas para marcar el significado de los datos de la web. Se caracteriza por no tener etiquetas predefinidas, sólo cuenta con un conjunto de reglas sintácticas que son útiles para crear lenguajes personalizados basados en XML. En la Figura 3.5 se detallan los componentes que contiene un documento XML.

3.6.2. RDF (Resource Description Framework)

RDF Es un lenguaje de marcado que se basa en XML. Es utilizado para describir los recursos web mediante metadatos. A diferencia de XML, RDF provee además una estructura de metadatos con significado que puede ser entendible por una máquina.

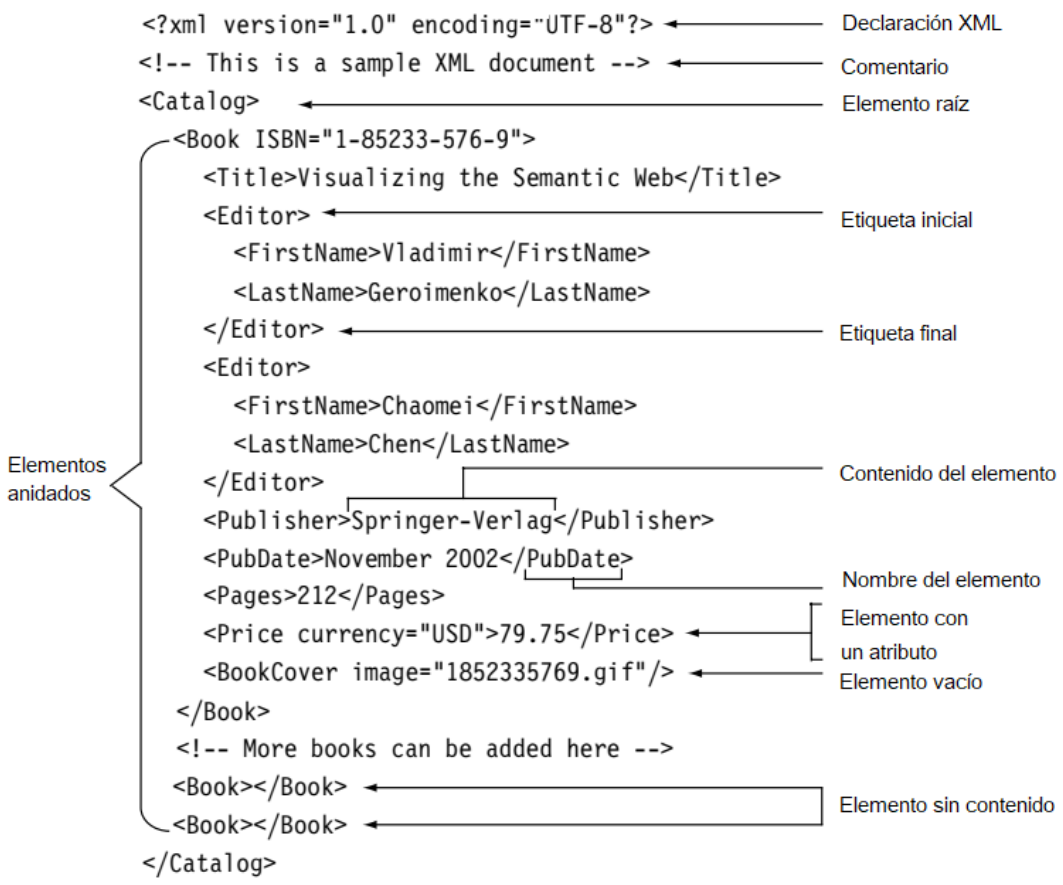

Figura 3.5: Anatomía de un documento XML [13]

RDF Schema

RDF Schema es una extensión de RDF que provee un vocabulario para expresar clases junto con sus relaciones, así como también para definir las propiedades de las clases. En la figura 3.6 se muestra un ejemplo sencillo con los componentes principales de un documento de RDF Schema.

3.6.3. OWL (Web Ontology Language)

OWL es un lenguaje de marcado basado en XML, RDF y RDF Schema que se utiliza para definir e instanciar ontologías web. OWL permite representar de forma

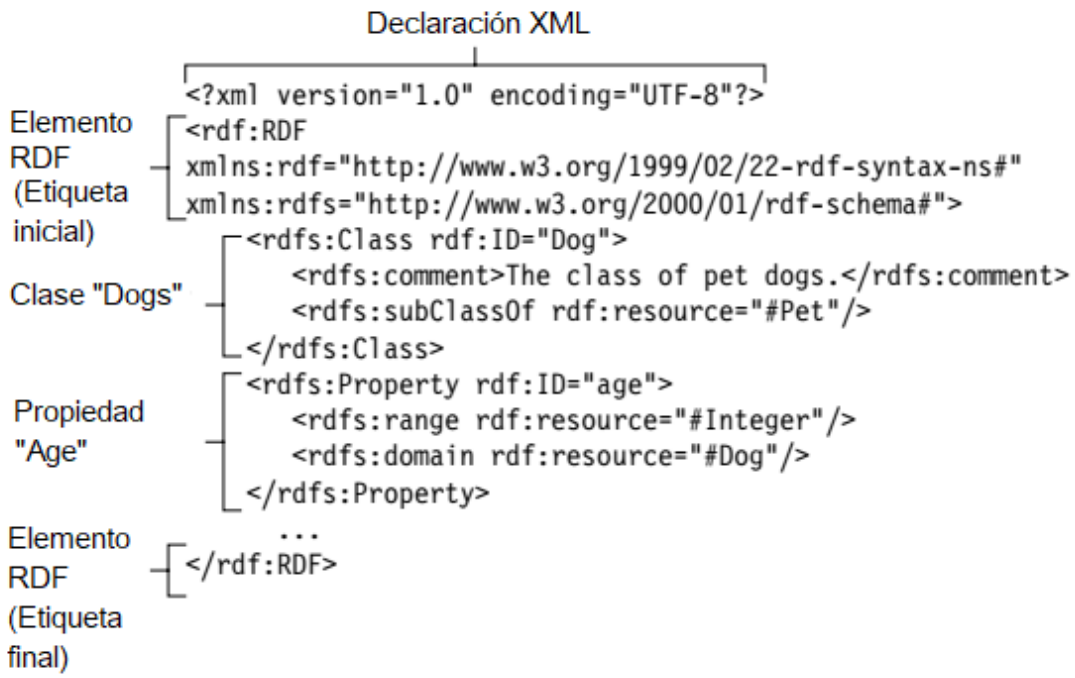

Figura 3.6: Ejemplo de un documento RDF Schema [13]

explícita el significado de términos y sus relaciones. Además proporciona facilidades para expresar la semántica de los datos web con opciones avanzadas de inferencia a comparación de los lenguajes en los que está basado OWL. En la Figura 3.7 se muestra un ejemplo de la estructura básica de un documento OWL.

3.6.4. Prótegé

Prótegé [28] es una plataforma libre y de código abierto con la que se puede crear y editar ontologías mediante una interfaz de usuario personalizable[1]. Fue desarrollado en la Universidad de Stanford en la década de 1980 y a lo largo de los años ha ayudado a los desarrolladores tanto a construir ontologías reutilizables como a construir sistemas basados en el conocimiento. Prótegé tiene dos versiones: la primera es un sistema de escritorio llamado Prótegé (ver Figura 3.8)que admite funciones avanzadas para la construcción y administración de ontologías en formato OWL.

[1]https://protege.stanford.edu/

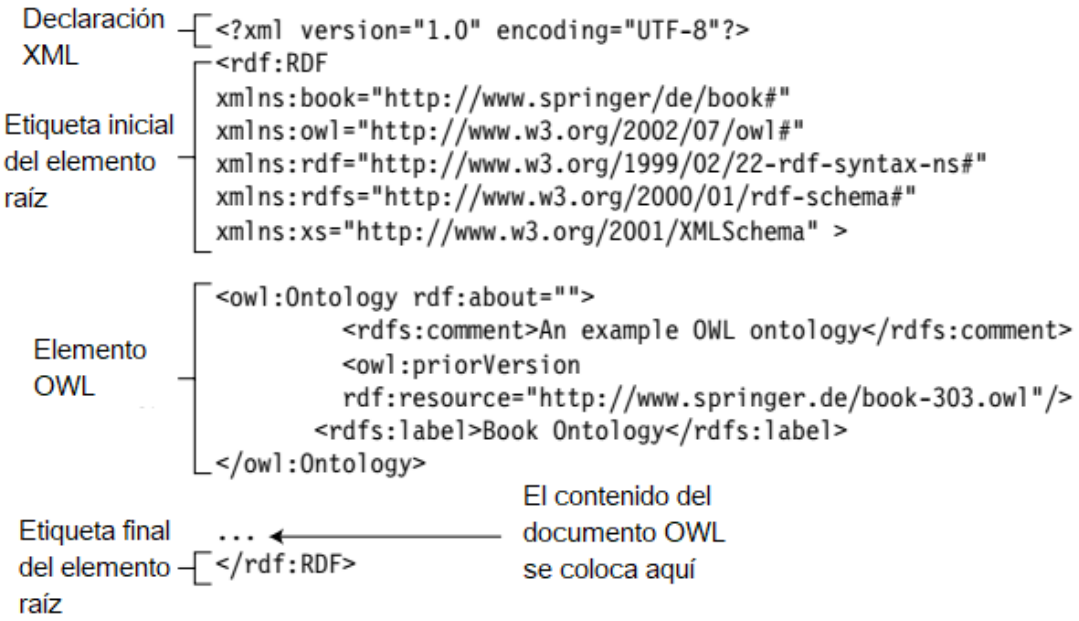

Figura 3.7: Estructura de un documento OWL [13]

La segunda es una versión web llamada WebProtégé (ver Figura 3.9) que ofrece un acceso mediante cualquier navegador con un diseño más sencillo de usar para tareas de ingeniería ontológica.

3.6.5. Sematch

Sematch[64][65] es un *framework* desarrollado por Ganggao Zhu del Grupo de Sistemas Inteligentes de la Universidad Politécnica de Madrid. Fue escrito en Python y proporciona herramientas de similitud que permiten calcular la similitud semántica de conceptos, palabras y entidades en grafos de conocimiento y se puede acceder al código fuente en GitHub[2]. Las medidas de similitud que utiliza son las medidas basadas en conocimiento, tanto las medidas basadas en el conocimiento estructural que toman en cuenta la profundidad, la distancia, el subsumidor menos común, etc., como también utiliza las medidas basadas en contenido de información. Las medidas implementadas son: Path, Wu & Palmer,Li, Resnik, Lin, Jiang & Conrad y una

[2]https://github.com/gsi-upm/sematch

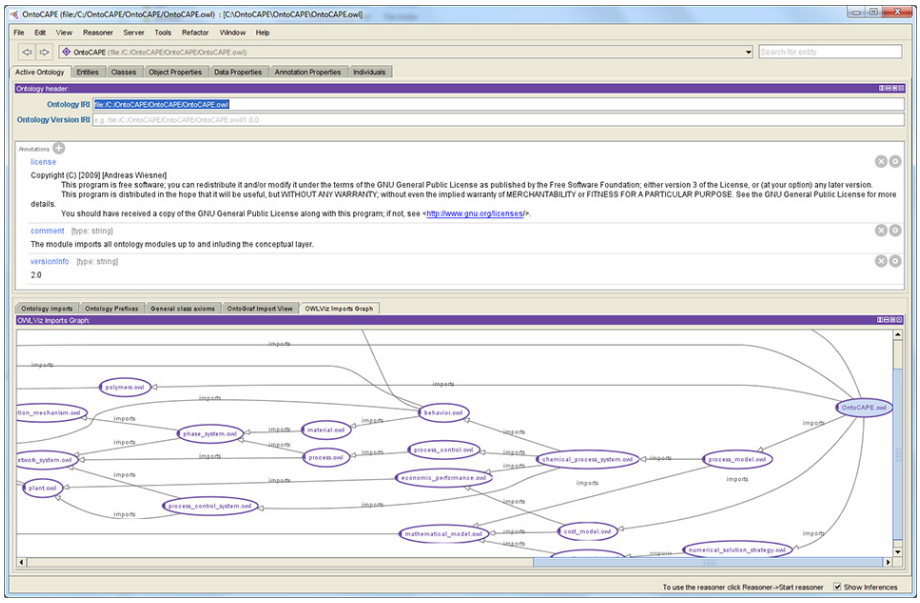

Figura 3.8: Interfaz de Prótegé versión escritorio [28]

medida propuesta por los autores llamada WPath[64]. Este *framework* funciona con conceptos de los grafos de conocimiento YAGO y DBpedia, existe una demostración web[3] del funcionamiento de este sistema donde se pueden hacer pruebas de las principales funciones del *framework* como se puede observar en la Figura 3.10. Este *framework* también puede ser adaptado para ontologías en formato OWL o RDF, este procedimiento fue realizado para poder utilizar las medidas implementadas en el *framework* en el sistema propuesto en esta tesis.

3.6.6. SML (Semantic Measures Library)

SML[4] es una biblioteca de código abierto desarrollado en Java por Sébastien Harispe[18] del LGI2P laboratory de la Ecole des mines d'Alès. Su objetivo es el cálculo y análisis de medidas semánticas como la similitud semántica, relación semántica, distancia semántica, etc., en entidades semánticas como conceptos e instancias. Dado que la biblioteca es genérica puede ser utilizada en diferentes ontologías como en las ontologías RDF, las ontologías OWL, WordNet, MesH, las ontologías

[3]http://sematch.cluster.gsi.dit.upm.es/
[4]http://www.semantic-measures-library.org/

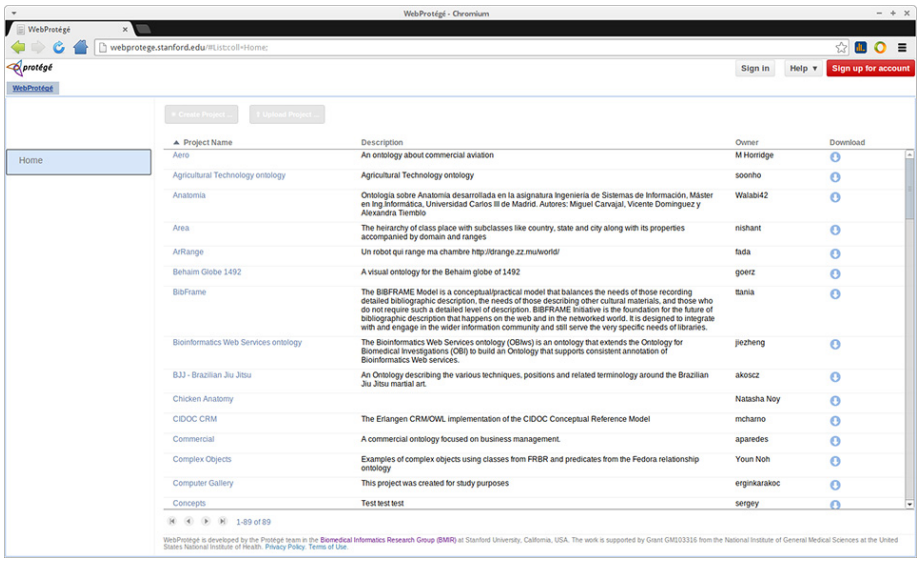

Figura 3.9: Interfaz de Prótegé versión web [28]

de genes (como OBO), SNOMED-CT, etc. Tiene dos versiones, SML Library que puede ser obtenida desde GitHub[5], importado desde el repositorio de Maven[6] o descargado en un archivo jar[7] de Java. La segunda versión es SML Toolkit[8], que es un programa amigable con el usuario que no tiene conocimientos en programación con las funciones de SML Library.

[5]https://github.com/sharispe/slib/tree/master/slib-sml
[6]https://mvnrepository.com/artifact/com.github.sharispe/slib-sml
[7]http://www.semantic-measures-library.org/sml/downloads/releases/sml/0.9.4/slib-dist-0.9.4d-all-jar.jar
[8]http://www.semantic-measures-library.org/sml/index.php?q=toolkit

Figura 3.10: Demostración web del *framework* Sematch [64][65]

Capítulo 4

Diseño de la propuesta de solución

Al tener las bases necesarias con los conceptos teóricos para la realización de esta investigación presentados en el Capítulo 3 y además de un panorama de los trabajos relacionados en el estado del arte 2. En este capítulo se propone un algoritmo para la evaluación de relaciones taxonómicas en ontologías de dominio.

4.1. Algoritmo propuesto

El algoritmo propuesto para evaluar las relaciones taxonómicas en ontologías de dominio se presenta de forma gráfica en la Figura 4.1. donde se resaltan las principales fases que lo componen, como se puede observar, en primer lugar se encuentra el preprocesamiento donde a una ontología en formato OWL se extraen los pares de conceptos de las relaciones taxonómicas, es decir, aquellos con una relación *subClassOf*. La siguiente fase es el cálculo de la similitud semántica que se realiza con la ayuda de las bibliotecas *Sematch* y *SML*. A continuación, se encuentra el cálculo de umbrales que se realiza a partir de los resultados obtenidos de similitud semántica y que servirán para la evaluación de las relaciones taxonómicas. Teniendo los resultados de similitud semántica, los umbrales y la anotación del experto (en caso que la ontología cuente con ello) se procede con la etapa de evaluación, donde se realiza la comparación de los resultados de similitud semántica con los umbrales para verificar si las relaciones taxonómicas son correctas. La anotación de un experto es una evaluación hecha por un experto en el dominio donde determina si el par de

conceptos forman una relación taxonómica correcta consultando un diccionario de hiperónimo-hipónimo y de la base de datos léxico-conceptual WordNet. Cuando se cuenta con esta anotación, se procede a realizar una comparación de los resultados obtenidos de la comparación anterior con los umbrales para verificar la efectividad del algoritmo. Al realizar las comparaciones antes mencionadas, el algoritmo tiene como salida la exactitud de las relaciones taxonómicas de la ontología, así como la exactitud del algoritmo al realizar la evaluación. Todos estos pasos se modelan en el Algoritmo 1 cuya entrada son dos listas con los conceptos que integran las relaciones taxonómicas existentes en la ontología, una correspondiente a los conceptos y la otra lista a los subconceptos (conceptos[], subconcepto[]) y $|RT|$ que es el número total de relaciones taxonómicas en la ontología. Hay que tomar en cuenta que este algoritmo se aplica una vez para cada tipo de medida utilizada en esta tesis, es decir, se aplica para las medidas de similitud basadas en estructura, y se vuelve a aplicar para las medidas de similitud basadas en características y así sucesivamente. A continuación, se describe cada paso del algoritmo.

Algoritmo 1 Algoritmo propuesto

Entrada: conceptos[], subconcepto[] experto[], $|RT|$
Salida: Exactitud de cada medida de similitud semántica

1: {Llamado Sematch o SML para usar las funciones de similitud semántica.}
2: **para** i=0 hasta $|RT|$ **hacer**
3: res_{medida}[i] $\leftarrow$ sim_{medida}(concepto[i], subconcepto[i])
4: **fin para**
5: $umb_{medida} \leftarrow$ promedio(res_{medida}[])
6: $umb_{promedio} \leftarrow$ promedio(umb_{medida})
7: $tablaVerdad_{medida}$[] $\leftarrow$ compUmbral(res_{medida}[], umb_{medida}, $|RT|$)
8: $expertoTabla_{medida} \leftarrow$ compExperto($tablaVerdad_{medida}$[], experto[],$|RT|$)
9: ex_{medida}[] $\leftarrow$ Exactitud($tablaVerdad_{medida}$[])
10: $tablaPromGeneral$[] $\leftarrow$ promedioSist(todos los res_{medida},$umbral_{promedio}$,$|RT|$)
11: $PG \leftarrow$ Exactitud($tablaPG$[])
12: $expertoPromTabla$[] $\leftarrow$ compExperto($tablaPG$[], experto[],$|RT|$)
13: $PromExperto \leftarrow$ Exactitud($expertoPromTabla$[])

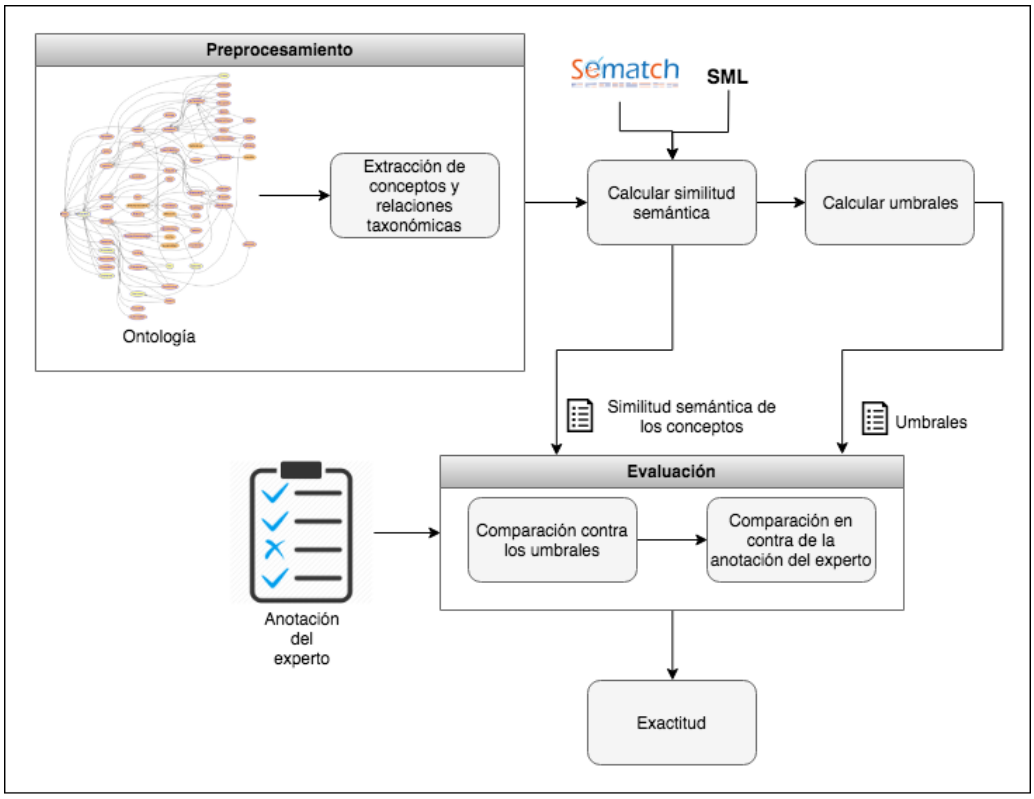

Figura 4.1: Algoritmo para evaluar las relaciones taxonómicas

4.1.1. Preprocesamiento

En la fase de preprocesamiento, se extraen los conceptos y las relaciones taxonómicas de la ontología utilizando SPARQL[1] y Apache Jena[2]. Esta fase es importante para definir la entrada de datos del Algoritmo 1.

4.1.2. Cálculo de la similitud semántica

Esta fase corresponde a los pasos 1-4 del Algoritmo 1. Para el funcionamiento de esta fase, necesitan de dos recursos: el *framework* desarrollado en Python llamado Sematch[3] que se utiliza para calcular la similitud semántica con las medidas basadas en estructura y en características, el segundo recurso es una biblioteca llamada *Semantic Measures Library* (SML)[4] que fue desarrollada en Java que se utiliza para calcular la similitud semántica con las medidas basadas en contenido de información. Teniendo estos dos recursos listos, se procede a calcular la similitud semántica de los conceptos que integran la relación semántica con la lista obtenida en la fase de preprocesamiento. Se obtienen los resultados de similitud para cada medida de similitud utilizada en esta investigación (ver sección 3.5.3): las medidas basadas en estructura (Path, WUP , Li), las medidas basadas en características (CMatch, RE, San), las medidas basadas en contenido de información (Resnik, Jiang-Conrath, Mazandu-Mulder) y las medidas híbridas (Zhou, HSS3 y HSS4). Los resultados obtenidos en esta fase se guardan en una lista por cada medida de similitud semántica ($res_{medida}[]$).

4.1.3. Cálculo de los umbrales

Para esta fase, se siguen los pasos 5 y 6 del Algoritmo 1. Teniendo los resultados de similitud de los conceptos de las relaciones taxonómicas de la ontología, se procede a hacer el cálculo de los umbrales siguiendo la Ecuación 4.1. Primero se calculan los umbrales (umb_{medida}) desde los resultados de cada medida de similitud ($res_{medida}[]$). Después se procede a calcular un umbral promedio ($umb_{promedio}$),

[1] https://www.w3.org/TR/rdf-sparql-query/
[2] https://jena.apache.org
[3] https://github.com/gsi-upm/sematch
[4] http://www.semantic-measures-library.org/sml/

siguiendo la ecuación 4.2 por cada tipo de medida de similitud.

$$umb_{sim_i} = \frac{\sum res_{sim_i}(u, v)}{RT} \tag{4.1}$$

El umbral promedio se obtiene al aplicar la ecuación 4.2, donde N es el total de las medidas de la categoría utilizada, por ejemplo, si se están usando las medidas de similitud basadas en estructura, $N = 3$ que son las tres medidas basadas en estructura usadas en esta tesis.

$$umb_{promedio} = \frac{\sum umb_{sim_i}}{N} \tag{4.2}$$

4.1.4. Evaluación

Al tener los resultados de las fases previas de Cálculo de similitud semántica 4.1.2 y Cálculo de los umbrales 4.1.3, se procede a realizar la evaluación de las relaciones taxonómicas. En el Algoritmo 1 esta fase corresponde a los pasos 7 y 8. Esta subsección se divide en dos pasos:

1. Comparación contra los umbrales promedio.
2. Comparación contra la anotación del experto.

El paso 2 tiene como entrada una lista con la anotación de un experto y solo se aplica a aquellas ontologías que cuentan con una anotación de un experto en el dominio, como es el caso de la ontología de Inteligencia Artificial [66] . Para las demás ontologías, solo se realiza la parte uno.

Comparación contra los umbrales promedio

Para este paso, se utiliza la función del Algoritmo 2. Esta función toma como entrada la lista de resultados de similitud de las medidas de similitud ($res_{medida}[]$), el umbral correspondiente a esa medida (umb_{medida}), y el total de relaciones taxonómicas en la ontología ($|RT|$). Después se realiza la comparación contra el umbral, si el resultado de similitud res_{medida}[i] es mayor o igual al umb_{medida} se toma como verdadera (1) la relación taxonómica, de lo contrario se considera falsa (0). Como salida, retorna una lista ($tablaVerdad_{medida}[]$) donde se guardan los resultados obtenidos en esta comparación.

Algoritmo 2 Función *compUmbral*

Entrada: $res_{medida}[]$, umb_{medida}, $|RT|$
Salida: $tablaVerdad_{medida}[]$

1: **para** i=0 hasta $|RT|$ **hacer**
2: **si** res_{medida}[i] >= umb_{medida} **entonces**
3: $tablaVerdad_{medida}$[i] $\leftarrow$ 1
4: **si no**
5: $tablaVerdad_{medida}$[i] $\leftarrow$ 0
6: **fin si**
7: **fin para**
8: **devolver** $tablaVerdad_{medida}[]$

Comparación contra la anotación del experto

Para este paso se utiliza la función del Algoritmo 3. Esta función toma como entrada la lista $tablaVerdad_{medida}[]$ obtenida en el paso anterior, la lista anotada del experto ($experto[]$) y el total de relaciones taxonómicas en la ontología ($|RT|$). Se realiza la comparación de la siguiente manera: si el resultado de la relación taxonómica en $tablaVerdad_{medida}$[i] fue verdadera (1) y la anotación del experto que también fue verdadera (1) entonces esta coincidencia se toma como verdadera y se guarda en $expertoTabla_{medida}[]$. Lo mismo sucede cuando la relación taxonómica fue considerada falsa y la anotación del experto coincide en tomarla como falsa, esta comparación es tomada como verdadera. De otra manera, la comparación es considerada falsa.

Algoritmo 3 Función *compExperto*

Entrada: $tablaVerdad_{medida}[]$, experto[], ,$|RT|$
Salida: $expertoTabla_{medida}$

1: **para** i=0 hasta $|RT|$ **hacer**
2: **si** $tablaVerdad_{medida}$[i] = 1 **and** experto[i] >= 1 **entonces**
3: $expertoTabla_{medida}$[i] $\leftarrow$ 1
4: **si no**, **si** $tablaVerdad_{medida}$[i] = 0 **and** experto[i] = 0 **entonces**
5: $exactitudTabla_{medida}$[i] $\leftarrow$ 1
6: **si no**
7: $expertoTabla_{medida}$[i] $\leftarrow$ 0
8: **fin si**
9: **fin para**
10: **devolver** $expertoTabla_{medida}$

4.1.5. Exactitud

En esta fase, teniendo los resultados obtenidos en la fase de Evaluación 4.1.4, se calculan las exactitudes por medio de la ecuación de exactitud 4.3. Donde la *CantidadDeCasosCorrectos* se refiere a aquellas relaciones taxonómicas que se tomaron como verdaderas dependiendo del umbral tomado y *TotalDeCasos* es el total de relaciones taxonómicas en la ontología.

En el Algoritmo 1, esta fase corresponde a los pasos 9-13.

$$Exactitud = \frac{CantidadDeCasosCorrectos}{TotalDeCasos} \tag{4.3}$$

Con la ecuación 4.3 se obtienen las siguientes exactitudes:

- Exactitudes para cada medida de similitud semántica.
- Exactitud promedio del algoritmo.
- Exactitud promedio de la comparación contra el experto.

Exactitudes para cada medida de similitud semántica

Esta parte corresponde al paso 9 del Algoritmo 1. Sólo se procede a calcular la exactitud mediante la ecuación 4.3 con los resultados de cada medida de similitud en $tablaVerdad_{medida}[]$.

Exactitud promedio del algoritmo

Para esta parte, primeramente se realiza otra comparación pero tomando en cuenta todos los $res_{medida}[]$ calculados contra el umbral promedio utilizando la función del Algoritmo 4. Esta función toma como entrada todos los $res_{medida}[]$, el $umb_{promedio}$ y el total de relaciones taxonómicas $|RT|$. El procedimiento es el mismo como en la función 2 a diferencia que en esta función se calcula el promedio de los resultados de todos los $res_{medida}[]$ y este promedio es comparado contra el $umb_{promedio}$. Los resultados de la comparación son guardados en $tablaPG[]$. Esta parte corresponde al paso 10 del Algoritmo 1.

Teniendo la lista $tablaPG[]$, se procede a calcular la exactitud promedio del algoritmo, que se refiere a qué tan eficiente fue el algoritmo al realizar la evaluación

Algoritmo 4 Función *promedioSist*

Entrada: todos los $res_{medida}[]$, $umb_{promedio}$, $|RT|$
Salida: $tablaPG[]$

1: **para** i=0 hasta $|RT|$ **hacer**
2: **si** promedio(todos los $res_{medida}[]$)>= $umb_{promedio}$ **entonces**
3: $tablaPG$[i] $\leftarrow$ 1
4: **si no**
5: $tablaPG$[i] $\leftarrow$ 0
6: **fin si**
7: **fin para**
8: **devolver** $tablaPG[]$

de las relaciones taxonómicas de la ontología. Esta exactitud es calculada aplicando la ecuación 4.3 a la $tablaPG[]$. Esta parte corresponde al paso 11 del Algoritmo 1.

Exactitud promedio de la comparación contra el experto

Como se dijo anteriormente, este paso es exclusivo para aquellas ontologías que cuentan con una lista con la anotación de un experto en el dominio. Esta parte corresponde los pasos 12 y 13 del Algoritmo 1. Primero se hace una comparación de los resultados de la lista $tablaPG[]$ con la lista anotada del experto, de la misma manera que se realizó cuando se hizo la comparación de los resultados de similitud contra la lista anotada del experto en 4.1.4 pero en esta ocasión en vez de usar la lista $tablaVerdad_{medida}[]$, se usó la lista $tablaPG[]$. Al final se obtiene otra lista $expertoPromTabla[]$ con los resultados de esta comparación en particular. Esta parte corresponde al paso 12 del Algoritmo 1.

Para obtener la exactitud promedio de la comparación contra el experto, se aplica la ecuación 4.3 a la lista $expertoPromTabla[]$. Esta parte corresponde al paso 13 del Algoritmo 1.

En el siguiente capítulo se presentan los resultados experimentales de la aplicación del Algoritmo 1 propuesto en esta sección con ontologías de diferentes dominios.

Capítulo 5

Resultados experimentales

En este capítulo se muestran los resultados que se obtuvieron al aplicar el Algoritmo 1 donde se evalúan las relaciones taxonómicas en diferentes ontologías (IA,SCORM y EOR), principalmente del área de biomedicina y sensores. Este capítulo se divide en cuatro secciones: en la primera sección se presentan el conjunto de datos con los que se trabajaron. En la segunda sección se presentan las evaluaciones hechas con las ontologías que cuentan con la anotación de un experto. En la tercera sección, se presentan las evaluaciones hechas para las ontologías del dominio de biomedicina. En la cuarta sección, se detallan las evaluaciones hechas para las ontologías en el dominio de sensores.

5.1. Conjuntos de datos

En esta sección se presentan los conjuntos de datos utilizados para el desarrollo del experimento de evaluación de relaciones taxonómicas de ontologías de dominio utilizando las medidas de similitud semántica (ver Sección 3.5.3) y así lograr los objetivos específicos presentados en la sección 1.2.

5.1.1. Ontologías con anotación de expertos

El primer conjunto de datos considerados para esta investigación, está constituido por tres ontologías de dominio: Inteligencia Artificial (IA) [66] , Sharable Content Object Reference Model (SCORM)[66] y Enhanced Oil Recovery (EOR)[30]. Estas

tres ontologías cuentan con la anotación de un experto para cada relación taxonómica. En la Tabla 5.1 se muestra el total de conceptos y las relaciones taxonómicas de cada ontología.

Tabla 5.1: Conjunto de datos con anotaciones de expertos

Ontología	Conceptos	Relaciones taxonómicas
IA	277	205
SCORM	1461	1038
EOR	83	80

En la Tabla 5.2 se presenta el total de relaciones taxonómicas que fueron revisadas por los expertos, para la ontología de IA, todas las relaciones fueron revisadas, en cambio para las ontologías de SCORM y EOR solo se tomó una muestra de las relaciones taxonómicas para ser revisadas.

Tabla 5.2: Total de relaciones taxonómicas revisadas por los expertos

Ontología	Relaciones taxonómicas	Relaciones taxonómicas revisadas
IA	205	205
SCORM	1038	100
EOR	83	35

Para cada ontología existen tres anotaciones de tres expertos diferentes como se muestra en la tabla 5.3.

Tabla 5.3: Expertos para cada ontología

Ontología	Expertos
IA	Experto 1, Experto 2, Experto 3
SCORM	Experto 2, Experto 2, Experto 3
EOR	Experto 3, Experto 2, Experto 3

5.1.2. Ontologías del dominio de biomedicina

El segundo conjunto de datos contiene cinco ontologías del área de biomedicina usadas para la investigación de Martínez-Romero et al. [26].

En esta investigación, se desarrolló el sistema Ontology Recommender 2.0 que es una nueva versión del sistema de recomendación de ontologías de NCBO. Este sistema toma en cuenta cuatro diferentes criterios: la medida en que la ontología cubre los datos de entrada, la aceptación de la ontología en la comunidad de biomedicina, el nivel de detalle que las clases de la ontología cubren los datos de entrada y la especialización de la ontología para el dominio de los datos de entrada. A continuación se describen brevemente las ontologías:

- NCIT[42]: *National Cancer Institute Thesaurus*, es un vocabulario que incluye una amplia cobertura del dominio del cáncer, donde se incluyen las enfermedades relacionadas con el cáncer, los hallazgos y las anomalías, desarrollada por el Centro de Informática Biomédica y Tecnología de la Información del NCI.

- OBI[3]: *Ontology for Biomedical Investigations*, es una ontología que proporciona términos que describen los aspectos sobre la realización de las investigaciones en los dominios biológicos y médicos, OBI cubre todas las fases del proceso de investigación.

- SYN[24]: *Sage Bionetworks Synapse Ontology*, es un vocabulario que integra el fenotipo clínico y los datos genómicos junto con una capa molecular intermedia.

- CL[4]: *Cell Ontology*, es una ontología desarrollada para las bases de datos de bioinformática que necesita de un vocabulario controlado de tipos de células. Esta ontología cubre tipos de células desde procariotas hasta mamíferos, excluyendo los tipos de células vegetales.

- SYMP[1]: *Symptom Ontology*, es una ontología que sirve para capturar y documentar de una manera sólida los signos y Síntomas. Fue desarrollado por Lynn Schriml.

Las ontologías mencionadas anteriormente se clasifican en diferentes dominios dependiendo del propósito por el cual fueron diseñadas, la clasificación se muestra a continuación:

- Investigaciones biomédicas: NCIT, OBI.

[1] https://bioportal.bioontology.org/ontologies/SYMP

- Signos y Síntomas: NCIT, SYMP.
- Tipos de células: SYN, CL.

Las ontologías mostradas en la Tabla 5.4 fueron analizadas por el sistema Ontology Recommender versión 1.0 y con la versión 2.0 propuesta en [26]. La ontologías recomendadas son en base al tipo de dominio de entrada que se presentan en la Tabla 5.5.

Tabla 5.4: Conjunto de datos del dominio de biomedicina

Ontología	Conceptos	Relaciones taxonómicas
NCIT	138,291	135,348
OBI	3,398	3,569
SYN	14,462	15,352
CL	6,635	3,243
SYMP	944	848

Estas ontologías también fueron analizadas con el algoritmo propuesto en esta investigación, evaluando sus relaciones taxonómicas y donde se compara con los resultados de los sistemas de recomendación con las exactitudes obtenidas del algoritmo propuesto.

En la tabla 5.4 se muestran las ontologías a evaluar y el total de conceptos y relaciones taxonómicas con las que cuenta cada una.

Tabla 5.5: Dominio de entrada y resultados de los sistemas de recomendación, para el dominio de biomedicina

Dominio	Resultado v1.0	Resultado v2.0
Investigaciones biomédicas	NCIT	OBI
Signos y Síntomas	NCIT	SYMP
Tipos de celulas	SYN	CL

5.1.3. Ontologías del dominio de sensores

El tercer conjunto de datos son cuatro ontologías del dominio de sensores, un breve resumen de las características de cada ontología se presenta a continuación:

- MMI Device Ontology[39]: es una ontología desarrollada para dispositivos marinos debido a la falta de un vocabulario común, a la dificultad de encontrar información relacionada con un dispositivo y a los enfoques no coordinados en la gestión de datos de sensores.

- OntoSensor[19]: esta ontología fue creada para construir un respositorio de conocimiento de sensores para consulta e inferencia. Incluye definiciones de conceptos y propiedades de SensorML y IEEE SUMO. Es una ontología de gran peso y tiene una complejidad de uso alta pero carece de la capacidad de describir las observaciones del sensor.

- CSIRO Sensor Ontology[29]: es una ontología genérica para describir sensores y sus capacidades. Se diseñó para ser usada en la integración, búsqueda y clasificación de datos. Además proporciona una descripción semántica de los sensores para su uso en flujos de trabajo.

- Intelligent Environment: es un modelo ontológico para la representación semántica de objetos, personas e interacciones que existen en ambientes académicos. Además proporciona soporte para el razonamiento inteligente e inferencias. Esta ontología está formada por un conjunto de ontologías modulares que se describen a continuación:

    - Person Ontology: representa todos los posibles usuarios de la institución académica. Se define un identificador único para cada tipo de persona.
    - Physical Space: representa cualquier tipo de ubicación física.
    - Sensor Network: representa una colección de sensores físico y móviles, además de actuadores. Su objetivo es facilitar la representación de los datos y razonamiento sobre el contexto físico y del usuario.

En la Tabla 5.6 se muestran el total de clases o conceptos, el número de relaciones taxonómicas (RT) y la profundidad máxima de cada ontología del dominio de sensores usadas en esta investigación.

Tabla 5.6: Ontologías de sensores

Ontología	Clases	RT	Profundidad
Intelligent Environment	81	27	4
Onto Sensor	289	252	12
Sensor Ontology	71	64	5
Device Ontology	56	40	5

5.2. Evaluación con anotaciones de expertos

En esta sección, se evalúan las relaciones taxonómicas de ontologías de dominio que cuentan con la anotación de un experto que evalúa cada relación taxonómica y verifica si las relaciones son correctas. Cada ontología de este conjunto de datos tiene tres anotaciones de expertos diferentes. Las relaciones taxonómicas fueron evaluadas con todas las medidas basadas en conocimiento que fueron presentadas en el Capítulo 4.

5.2.1. Resultados de las medidas basadas en estructura

En esta sección se presentan los resultados obtenidos usando el Algoritmo 1 con las medidas basadas en estructura, con el conjunto de datos que tienen las anotaciones de los expertos.

Inteligencia Artificial (IA)

Para la ontología de Inteligencia Artificial, se obtuvieron los umbrales (Umb_{sim_i}, i representa cada medida de similitud semántica) de las medidas basadas en estructura (ver Tabla 5.7) que se calculan siguiendo la Ecuación 4.1. A partir de los tres umbrales, se calcula el umbral promedio ($Umb_{promedio}$) por medio de la Ecuación 4.2.

Tabla 5.7: Umbrales de las medidas basadas en estructura en la ontología IA

$Umb_{sim_{Path}}$	$Umb_{sim_{WUP}}$	$Umb_{sim_{Li}}$	$Umb_{promedio}$
0.500	0.778	0.680	0.652

Con los umbrales de la tabla 5.7 se procede a calcular las exactitudes para cada medida de similitud (Ex_{sim}), comparando los valores de similitud obtenidos con los

valores de la anotación de cada experto. Las exactitudes obtenidas se muestran en la Tabla 5.8, además de la exactitud promedio del experto (PE) y el promedio general del algoritmo (PG).

Tabla 5.8: Exactitudes de las medidas basadas en estructura para la ontología IA

Experto	$Ex_{sim_{Path}}$	$Ex_{sim_{WUP}}$	$Ex_{sim_{Path}}$	PG	PE
1	0.888	0.912	0.907	0.927	0.912
2	0.922	0.927	0.922	0.927	0.927
3	0.976	0.980	0.980	0.927	0.980
$PE_{promedio}$					0.940

En general, el algoritmo obtiene una exactitud promedio general (PG) de 92.7 % para los tres expertos. El PE del Experto 1 obtiene una diferencia del 1.5 % con respecto del PG. El PE del Experto 3 obtiene el mejor valor de exactitud con un 98 %. En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 94 % y el algoritmo obtiene una exactitud (PG) de 92.7 %, entonces el Algoritmo 1 usando las medidas basadas en estructura se acerca a la evaluación hecha por los expertos con un error absoluto de 1.3 %. La ontología de Inteligencia Artificial mediante el Algoritmo 1 y la evaluación hecha por los expertos indican que sus relaciones taxonómicas son correctas al menos en un 92 %.

Sharable Content Object Reference Model (SCORM)

A continuación se presentan los resultados obtenidos para la ontología SCORM con las medidas basadas en estructura. Los umbrales para cada medida (Umb_{sim_i}) que se obtienen con la Ecuación 4.1 y el umbral promedio ($Umb_{promedio}$), obtenido usando con la Ecuación 4.2 se presentan en la Tabla 5.9.

Tabla 5.9: Umbrales de las medidas basadas en estructura para la ontología SCORM

$Umb_{sim_{Path}}$	$Umb_{sim_{WUP}}$	$Umb_{sim_{Li}}$	$Umb_{promedio}$
0.500	0.772	0.668	0.647

Las exactitudes obtenidas para cada medida de similitud (Ex_{sim_i}) por cada anotación del experto se presentan en la Tabla 5.10 donde se observa que con la medida

de Wu & Palmer[62] ($Ex_{sim_{WUP}}$) y la medida de Li[21] ($Ex_{sim_{Li}}$), se obtienen valores iguales en la exactitud. En relación a los expertos, la exactitud promedio (PE) con mejor valor la tiene el Experto 3 con un 84 %.

Tabla 5.10: Exactitudes de las medidas basadas en estructura para la ontología SCORM

Experto	$Ex_{sim_{Path}}$	$Ex_{sim_{WUP}}$	$Ex_{sim_{Li}}$	PG	PE
1	0.670	0.680	0.680	0.930	0.680
2	0.680	0.710	0.710	0.930	0.710
3	0.820	0.840	0.840	0.930	0.840
$PE_{promedio}$					0.743

En general, el algoritmo obtiene un promedio (PG) de 93 % en la evaluación de las relaciones taxonómicas de esta ontología con las medidas basadas en estructura. La exactitud promedio del experto (PE) que se acerca más al juicio emitido por el Algoritmo 1 lo tiene el Experto 3 con un promedio del 84 % y una diferencia de 11 % con la exactitud promedio general del algoritmo (PG).

En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 74.3 % y el algoritmo obtiene una exactitud (PG) de 93 %. El Algoritmo 1 usando las medidas basadas en estructura obtiene una diferencia con la evaluación de los expertos de un 18.7 %. La ontología SCORM mediante el Algoritmo 1 indica que al menos un 93 % de sus relaciones taxonómicas son correctas, en cambio con la anotación de los expertos se tiene que al menos un 74 % de la muestra de las relaciones que se tomaron para esta ontologías son correctas.

Enhanced Oil Recovery (EOR)

A continuación se presentan los resultados obtenidos del algoritmo propuesto con las medidas basadas en estructura de la ontología EOR. En la Tabla 5.11 se presentan los umbrales para cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$).

Tabla 5.11: Umbrales de las medidas basadas en estructura en la ontología EOR

$Umb_{sim_{Path}}$	$Umb_{sim_{WUP}}$	$Umb_{sim_{Li}}$	$Umb_{promedio}$
0.500	0.883	0.795	0.726

En la Tabla 5.12 se presentan las exactitudes obtenidas por cada medida (Ex_{sim_i}), la exactitud promedio general del algoritmo (PG) y la exactitud promedio de los expertos (PE). Con los resultados obtenidos se observa que las medidas de Wu & Palmer[62] ($Ex_{sim_{WUP}}$) y la medida de Li[21] ($Ex_{sim_{Li}}$), obtienen valores de exactitud iguales. El experto con mejor valor de exactitud promedio (PE) es el Experto 1 con un 97.1 %. La exactitud promedio general del algoritmo (PG) fue de un 57.1 %. El experto que coincide mejor con el sistema fue el Experto 2 con un promedio de 68.6 % y una diferencia del 11.5 %.

Tabla 5.12: Exactitudes de las medidas basadas en estructura en la ontología EOR

Experto	$Ex_{sim_{Path}}$	$Ex_{sim_{WUP}}$	$Ex_{sim_{Li}}$	PG	PE
1	0.943	0.971	0.971	0.571	0.971
2	0.600	0.686	0.686	0.571	0.686
3	0.743	0.771	0.771	0.571	0.771
$PE_{promedio}$					0.809

En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 80.9 % y el algoritmo obtiene una exactitud (PG) de 57.1 %. El Algoritmo 1 usando las medidas basadas en estructura tiene una diferencia con la evaluación de los expertos con un 23.8 %. La ontología EOR mediante el Algoritmo 1 indica que al menos un 57 % de sus relaciones taxonómicas son correctas, en cambio con la anotación de los expertos se tiene que al menos un 80 % de la muestra de las relaciones que se tomaron para esta ontologías son correctas.

5.2.2. Resultados de las medidas basadas en características

En esta sección se presentan los resultados de la evaluación de las relaciones taxonómicas de las tres ontologías: IA, SCORM y EOR con las medidas basadas en características.

Inteligencia Artificial (IA)

Los resultados de los umbrales para cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$), se presentan en la Tabla 5.13.

Tabla 5.13: Umbrales de las medidas basadas en características de la ontología IA

$Umb_{sim_{CMatch}}$	$Umb_{sim_{RE}}$	$Umb_{sim_{SAN}}$	$Umb_{promedio}$
0.500	0.792	0.685	0.659

Al observar los resultados de las exactitudes (Ex_{sim_i}) obtenidas para cada medida (ver Tabla 5.14), se tiene que la medida de Sánchez[47] ($Ex_{sim_{SAN}}$) tiene mejores valores de exactitud con los tres expertos. También se puede observar que el Experto 3 obtiene el mejor valor de exactitud promedio (PE) con un 97.6 %.

Tabla 5.14: Exactitudes de las medidas basadas en características para la ontología IA

Experto	$Ex_{sim_{CMatch}}$	$Ex_{sim_{RE}}$	$Ex_{sim_{SAN}}$	PG	PE
1	0.888	0.893	0.995	0.961	0.893
2	0.922	0.922	0.980	0.961	0.922
3	0.976	0.976	0.976	0.961	0.976
$PE_{promedio}$					0.930

El promedio general (PG) del algoritmo con la ontología de Inteligencia Artificial y las medidas basadas en características es de un 96.1 %. El algoritmo coincide más con las anotaciones del Experto 3 que tiene una exactitud promedio (PE) de 97.6 %, teniendo una diferencia de 1.5 %.

En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 93 % y el algoritmo obtiene una exactitud (PG) de 96.1 %. El Algoritmo 1 usando las medidas basadas en características tiene una diferencia con la evaluación de los expertos con un 3.1 %. La ontología de Inteligencia Artificial mediante el Algoritmo 1 y la evaluación hecha por los expertos indican que sus relaciones taxonómicas son correctas al menos en un 93 %.

Sharable Content Object Reference Model (SCORM)

Los umbrales para cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$) se presentan en la Tabla 5.15.

Con base en los resultados obtenidos de las exactitudes (ver Tabla 5.16),

se tiene que las medidas de Rodriguez & Egenhofer[38] ($Ex_{sim_{RE}}$) y la medida de

Tabla 5.15: Umbrales de las medidas basadas en características de la ontología SCORM

$Umb_{sim_{CMatch}}$	$Umb_{sim_{RE}}$	$Umb_{sim_{SAN}}$	$Umb_{promedio}$
0.500	0.774	0.668	0.647

Sánchez[47] ($Ex_{sim_{SAN}}$) tienen valores iguales de exactitud. En cuanto a los expertos, el que tiene mejor valor de exactitud promedio (PE) es el Experto 3 con un 82 %.

Tabla 5.16: Exactitudes de las medidas basadas en características para la ontología SCORM

Experto	$Ex_{sim_{CMatch}}$	$Ex_{sim_{RE}}$	$Ex_{sim_{SAN}}$	PG	PE
1	0.670	0.700	0.700	0.930	0.700
2	0.680	0.710	0.710	0.930	0.710
3	0.820	0.820	0.820	0.930	0.820
$PE_{promedio}$					0.743

En general el promedio del algoritmo (PG) es de un 93 %. El experto que mejor coincide con el algoritmo es el Experto 3 que tiene una exactitud promedio (PE) de 82 % con una diferencia de 11 %.

En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 74.3 % y el algoritmo obtiene una exactitud (PG) de 93 %. El Algoritmo 1 usando las medidas basadas en características tiene una diferencia con la evaluación de los expertos de un 18.7 %. La ontología SCORM mediante el Algoritmo 1 indica que al menos un 93 % de sus relaciones taxonómicas son correctas, en cambio con la anotación de los expertos se tiene que al menos un 74 % de la muestra de las relaciones que se tomaron para esta ontologías son correctas.

Enhanced Oil Recovery (EOR)

Para la ontología de EOR, se presentan los umbrales obtenidos para cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$) en la Tabla 5.17.

Con los valores obtenidos de exactitud (ver Tabla 5.18), se observa que

con las medidas de Rodriguez & Egenhofer[38] ($Ex_{sim_{RE}}$) y Sánchez[47] ($Ex_{sim_{SAN}}$) se obtiene valores iguales de exactitud. En cuanto a los expertos, el que obtuvo el mejor valor de exactitud promedio es el Experto 1 con un 97.1 %.

Tabla 5.17: Umbrales de las medida basadas en características de la ontología EOR

$Umb_{sim_{CMatch}}$	$Umb_{sim_{RE}}$	$Umb_{sim_{SAN}}$	$Umb_{promedio}$
0.495	0.878	0.791	0.721

Tabla 5.18: Exactitudes de las medidas basadas en estructura para la ontología EOR

Experto	$Ex_{sim_{CMatch}}$	$Ex_{sim_{RE}}$	$Ex_{sim_{SAN}}$	PG	PE
1	0.943	0.943	0.943	0.571	0.971
2	0.629	0.714	0.714	0.571	0.714
3	0.743	0.800	0.800	0.571	0.800
$PE_{promedio}$					0.828

El promedio general del algoritmo (PG) es de un 57.1 %. El experto que más se acerca a la exactitud promedio (PG) del algoritmo es el Experto 2 con una diferencia de 14.3 %. El algoritmo no tiene resultados favorables con esta ontología porque se obtiene un 57.1 % de exactitud utilizando las medidas basadas en características. Sin embargo, con la anotación del experto si se tienen resultados favorables por arriba del 80 %. Con una exactitud promedio (PE) de 97.1 % con el Experto 1 en la evaluación de las relaciones taxonómicas.

En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 82.8 % y el algoritmo obtiene una exactitud (PG) de 57.71 %. El Algoritmo 1 usando las medidas basadas en características tiene una diferencia con la evaluación de los expertos de un 25.7 %. La ontología EOR mediante el Algoritmo 1 indica que al menos un 57 % de sus relaciones taxonómicas son correctas, en cambio con la anotación de los expertos se tiene que al menos un 82.8 % de la muestra de las relaciones que se tomaron para esta ontologías son correctas.

5.2.3. Resultados de las medidas basadas en contenido de información

A continuación se presentan los resultados de la evaluación de las relaciones taxonómicas utilizando las medidas basadas en contenido de información en las tres ontologías: IA, SCORM y EOR.

Inteligencia Artificial (IA)

Los umbrales para cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$) se presentan en la Tabla 5.19.

Tabla 5.19: Umbrales de las medidas basadas en contenido de información en la ontología IA

$Umb_{sim_{Lim}}$	$Umb_{sim_{Maz}}$	$Umb_{sim_{JC}}$	$Umb_{sim_{Res}}$	$Umb_{promedio}$
0.804	0.677	0.868	0.555	0.726

Los valores obtenidos de exactitud para cada medida (Ex_{sim_i}), la exactitud promedio del algoritmo (PG) y la exactitud promedio del experto (PE) se muestran en la Tabla 5.20. Como se observa, se tiene que la medida de Resnik[37] ($Ex_{sim_{Res}}$) tiene los mejores valores de exactitud arriba del 95 %. El experto con mejor valor de exactitud promedio (PE) es el Experto 3 con un 98.5 %.

Tabla 5.20: Exactitudes de las medidas basadas en contenido de información para la ontología IA

Experto	$Ex_{sim_{Lin}}$	$Ex_{sim_{Maz}}$	$Ex_{sim_{JC}}$	$Ex_{sim_{Res}}$	PG	PE
1	0.932	0.932	0.932	0.951	0.473	0.927
2	0.971	0.971	0.971	0.980	0.473	0.971
3	0.985	0.985	0.985	0.985	0.473	0.985
$PE_{promedio}$						0.961

En general, el algoritmo tiene una exactitud promedio (PG) de un 47.3 %, un promedio por debajo del 50 %, por lo cual las medidas de similitud basadas en contenido de información no obtienen resultados favorables para esta ontología. Por otro lado, la exactitud promedio de los expertos (PE) obtiene valores mayores al 90 %.

En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 96.1 % y el algoritmo obtiene una exactitud (PG) de 47.3 %. El Algoritmo 1 usando las medidas basadas en estructura obtiene una diferencia con la evaluación de los expertos de un 48.8 %. La ontología IA mediante el Algoritmo 1 indica que al menos un 47 % de sus relaciones taxonómicas son correctas, en cambio con la anotación de los expertos se tiene que al menos un 96 % de la muestra de las relaciones que se tomaron para esta ontologías son correctas.

Sharable Content Object Reference Model (SCORM)

Para esta ontología con la biblioteca SML se produce un error al querer calcular el contenido de información, esto se debe a que en esta ontología hay un ciclo detectado en un camino, por lo tanto arroja un error diciendo que la ontología no es un grafo acíclico dirigido (DAG). En Harispe et al. [18] se menciona dos tipos de medidas dependiendo de los grafos a los que se aplican. El primer tipo son las medidas adaptadas a grafos semánticos compuestos por múltiples predicados (predicado se refiere a las relaciones semánticas que se establecen entre instancias o clases, por ejemplo, subClassOf.) que potencialmente inducen a ciclos. El segundo tipo son las medidas adaptadas a las taxonomías, es decir, a grafos semánticas acíclicos compuestos por un predicado único que induce a la transitividad. Dentro de este segundo tipo de medidas se encuentran las medidas de similitud semánticas utilizadas en este trabajo y que se encuentran en la biblioteca SML.

El camino que causa este problema es el siguiente: component_of_manifest es sublcase de component, component es subclase de asset, asset es subclase de content_model_component, content_model_component es subclase de component. Este camino se puede observar mejor en la Figura 5.1.

Motivo por el cual ya no se muestran los resultados experimentales de las evaluaciones taxonómicas en esta ontología.

Enhanced Oil Recovery (EOR)

A continuación se presentan los resultados obtenidos de la evaluación de las relaciones taxonómicas en la ontología EOR con las medidas de similitud basadas en contenido de información. Los umbrales (Umb_{sim_i}) para cada medida y el umbral promedio ($Umb_{promedio}$) se presentan en la Tabla 5.21.

Tabla 5.21: Umbrales de las medidas basadas en contenido de información en la ontología EOR

$Umb_{sim_{Li}}$	$Umb_{sim_{Maz}}$	$Umb_{sim_{JC}}$	$Umb_{sim_{Res}}$	$Umb_{promedio}$
0.982	0.810	0.917	0.697	0.829

Los valores obtenidos de exactitud para cada medida (Ex_{sim_i}), la exactitud promedio del algoritmo (PG) y la exactitud promedio del experto (PE) se presenta

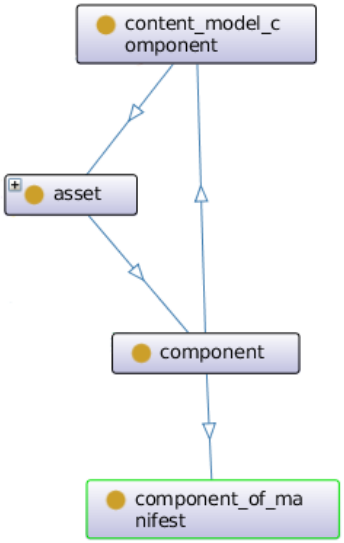

Figura 5.1: Camino en la ontología SCORM que causa conflicto.

en la Tabla 5.22. Con base en los resultados el Experto 1 tiene el mejor valor de exactitud promedio con un 97 %, las exactitudes obtenidas en cada medida son casi iguales en todos los resultados con una variación mínima en $Ex_{sim_{Res}}$.

Tabla 5.22: Exactitudes de las medidas basadas en contenido de información para la ontología EOR

Experto	$Ex_{sim_{Lin}}$	$Ex_{sim_{Maz}}$	$Ex_{sim_{JC}}$	$Ex_{sim_{Res}}$	PG	PE
1	0.971	0.971	0.971	0.971	0.514	0.970
2	0.714	0.714	0.714	0.714	0.514	0.743
3	0.800	0.800	0.800	0.829	0.514	0.829
$PE_{promedio}$						0.847

El promedio general del algoritmo (PG) es de un 51.4 %. El experto que más se acerca al algoritmo es el Experto 2, pero con una diferencia de 22.9 %. El algoritmo por si sólo, una vez más, no tiene resultados favorables con esta ontología teniendo un promedio general (PG) un poco mayor al 50 %, pero con las anotaciones de los expertos se tienen mejores resultados arriba del 70 % y llegando a un 97 %.

En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 84.7 % y el

algoritmo obtiene una exactitud (PG) de 51.4 %. El Algoritmo 1 usando las medidas basadas en contenido de información obtiene una diferencia con la evaluación de los expertos de un 33.3 %. La ontología EOR mediante el Algoritmo 1 indica que al menos un 51 % de sus relaciones taxonómicas son correctas, en cambio con la anotación de los expertos se tiene que al menos un 84 % de la muestra de las relaciones que se tomaron para esta ontologías son correctas.

5.2.4. Resultados de las medidas híbridas

En esta sección se presentan los resultados de la evaluación de las relaciones taxonómicas de las tres ontologías: IA, SCORM y EOR con las medidas híbridas.

Inteligencia Artificial (IA)

A continuación, se presentan los resultados obtenidos en la ontología IA con las medidas híbridas. Los umbrales de cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$) se muestran en la Tabla 5.23.

Tabla 5.23: Umbrales de cada medida híbrida en la ontología IA

$Umb_{sim_{Zhou}}$	$Umb_{sim_{HSS3}}$	$Umb_{sim_{HSS4}}$	$Umb_{promedio}$
0.184	0.085	0.132	0.134

La exactitud obtenida por cada medida (Ex_{sim_i}), la exactitud general promedio del algoritmo (PG) y la exactitud promedio del experto (PE) se presentan en la Tabla 5.24. Con base en los resultados obtenidos, se tiene que, con la medida HSS4 y HSS3 [32] se tienen los mejores valores de exactitud por arriba del 95 % para todos los expertos. El Experto 3 tiene el mejor valor de exactitud promedio (PE) obteniendo un 98.5 %.

En general, el algoritmo tiene un promedio (PG) del 28.3 %, estando por muy por debajo del 50 %. La exactitud promedio general del algoritmo (PG) no se acerca a los resultados de las exactitudes promedio de los expertos (PE) que están por arriba del 95 %.

En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 97.2 % y el algoritmo obtiene una exactitud (PG) de 28.3 %. El Algoritmo 1 usando las medidas

Tabla 5.24: Exactitudes de las medidas híbridas para la ontología IA

Experto	$Ex_{sim_{Zhou}}$	$Ex_{sim_{HSS3}}$	$Ex_{sim_{HSS4}}$	PG	PE
1	0.932	0.961	0.966	0.283	0.951
2	0.971	0.990	0.990	0.283	0.980
3	0.985	0.990	0.990	0.283	0.985
$PE_{promedio}$					0.972

híbrida tiene una diferencia con la evaluación de los expertos con un 68.9 %. La ontología IA mediante el Algoritmo 1 indica que al menos un 28 % de sus relaciones taxonómicas son correctas, en cambio con la anotación de los expertos se tiene que al menos un 97 % de la muestra de las relaciones que se tomaron para esta ontologías son correctas.

Sharable Content Object Reference Model (SCORM)

Para esta ontología con la biblioteca SML se produce un error al querer calcular el contenido de información, esto se debe a que en esta ontología hay un ciclo detectado en un camino, por lo tanto arroja un error diciendo que la ontología no es un grafo acíclico dirigido (DAG). En Harispe et al. [18] se menciona dos tipos de medidas dependiendo de los grafos a los que se aplican. El primer tipo son las medidas adaptadas a grafos semánticos compuestos por múltiples predicados (predicado se refiere a las relaciones semánticas que se establecen entre instancias o clases, por ejemplo, subClassOf.) que potencialmente inducen a ciclos. El segundo tipo son las medidas adaptadas a las taxonomías, es decir, a grafos semánticas acíclicos compuestos por un predicado único que induce a la transitividad. Dentro de este segundo tipo de medidas se encuentran las medidas de similitud semánticas utilizadas en este trabajo y que se encuentran en la biblioteca SML.

El camino que causa este problema es el siguiente: component_of_manifest es sublcase de component, component es sublcase de asset, asset es subclase de content_model_component, content_model_component es subclase de component. Este camino se puede observar mejor en la Figura 5.1.

Enhanced Oil Recovery (EOR)

A continuación, se presentan los resultados obtenidos de la evaluación de las relaciones taxonómicas de la ontología EOR utilizando las medidas híbridas. Los umbrales para cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$) se presentan en la Tabla 5.25.

Tabla 5.25: Umbrales de las medidas híbridas en la ontología EOR

$Umb_{sim_{Zhou}}$	$Umb_{sim_{HSS3}}$	$Umb_{sim_{HSS4}}$	$Umb_{promedio}$
0.189	0.159	0.243	0.197

Las exactitudes obtenidas de cada medida, la exactitud general promedio del algoritmo (PG) y la exactitud promedio del experto (PE) se muestran en la Tabla 5.26. Con los resultados obtenidos, se tiene que, con las medidas HSS4 y HSS3[32] tienen valores de exactitud iguales. El Experto 1 tiene el mejor valor de exactitud promedio (PE) con un 97.1 %.

Tabla 5.26: Exactitudes de la ontología EOR con las medidas híbridas

Experto	$Ex_{sim_{Zhou}}$	$Ex_{sim_{HSS3}}$	$Ex_{sim_{HSS4}}$	PG	PE
1	0.971	0.971	0.971	0.486	0.971
2	0.714	0.743	0.743	0.486	0.743
3	0.800	0.829	0.829	0.486	0.829
$PE_{promedio}$					0.848

La exactitud promedio general del algoritmo (PG) es de un 48.6 % un valor por debajo del 50 %, mientras que los resultados de exactitud promedio de los expertos (PE) obtienen valores arriba del 70 %.

En promedio, los tres expertos obtienen una exactitud ($PE_{promedio}$) de 84.8 % y el algoritmo obtiene una exactitud (PG) de 48.6 %. El Algoritmo 1 usando las medidas híbridas tiene una diferencia con la evaluación de los expertos con un 36.2 %. La ontología EOR mediante el Algoritmo 1 indica que al menos un 48.6 % de sus relaciones taxonómicas son correctas, en cambio con la anotación de los expertos se tiene que al menos un 84.8 % de la muestra de las relaciones que se tomaron para esta ontologías son correctas.

5.2.5. Resultados en todas las categorías de las medidas

En esta sección se muestra una recopilación de los resultados obtenidos de la evaluación de relaciones taxonómicas usando todas las medidas de similitud semántica utilizadas en este trabajo. Con los resultados obtenidos en este experimento sirven para comparar el comportamiento y efectividad que tiene el algoritmo propuesto con la anotación de expertos.

En la tabla 5.27 se muestran todas las exactitudes promedio por cada categoría de medidas de similitud semántica: las medidas basadas en estructura (Str), las medidas basadas en características (Ft), las medidas basadas en contenido de información (CI) y las medidas híbridas (H). En la última columna de la tabla se muestra un promedio de todas las exactitudes promedio por ontología ($PG_{promedio}$). Al final de la tabla, se muestra el promedio de las exactitudes (PG_i) por cada categoría de medida.

Tabla 5.27: Exactitud promedio para todas las categorías

Ontología	PG_{Str}	PG_{Ft}	PG_{CI}	PG_H	$PG_{promedio}$
IA	0.927	0.961	0.473	0.283	0.661
SCORM	0.930	0.930	-	-	0.465
EOR	0.571	0.571	0.514	0.486	0.536
Promedio	0.809	0.821	0.494	0.385	0.554

En la tabla 5.28 se muestran las exactitudes promedio de los expertos por cada categoría de medidas de similitud semántica y también el promedio de exactitud por cada ontología ($PG_{promedio}$). Al final se muestra el promedio por cada categoría de las exactitudes promedio.

Tabla 5.28: Exactitud para todas las categorías

Ontología	PE_{Str}	PE_{Ft}	PE_{CI}	PE_H	$PE_{promedio}$
IA	0.940	0.930	0.961	0.972	0.951
SCORM	0.743	0.743	-	-	0.743
EOR	0.809	0.828	0.847	0.848	0.833
Promedio	0.831	0.834	0.904	0.910	0.842

5.2.6. Discusión

Con base en los resultados de las tablas 5.27 y 5.28 se tiene que para la categoría de medidas basadas en estructura los expertos obtienen en promedio una exactitud de 83.1 % (ver columna PE_{Str} de la Tabla 5.28) y el Algoritmo 1 obtiene en promedio para esta categoría un 80.9 % (ver columna PG_{Ft} de la Tabla 5.27), siendo unos resultados muy cercanos indicando que las medidas basadas en estructura se comportan de manera eficiente al realizar la evaluación de las relaciones taxonómicas en las ontologías IA, SCORM y EOR teniendo una diferencia de 2.2 % entre los expertos y el algoritmo.

El Algoritmo 1 con las medidas basadas en características se obtiene un promedio de 82.1 % (ver columna PG_{Ft} de la Tabla 5.27) y los expertos obtienen en promedio una exactitud de 83.3 % (ver columna PE_{Ft} de la Tabla 5.28). Las medidas basadas en características, el Algoritmo 1 se comporta de manera eficiente en comparación a los expertos teniendo una diferencia de 1.2 % para las tres ontologías.

Para las medidas basadas en contenido de información se tomaron en cuenta solamente los resultados de las ontologías IA y EOR. El Algoritmo 1 obtiene un promedio de 49.4 % (ver columna PG_{CI} de la Tabla 5.27) y los expertos cuentan con un promedio de 90.4 % (ver columna PE_{CI} de la Tabla 5.28). Con esta categoría de medidas no se obtiene resultados favorables ya que la diferencia en promedios entre el algoritmo y los expertos es de 41 %.

Por último, para las medidas híbridas tomando en cuenta sólo las ontologías IA y EOR se tiene que el Algoritmo 1 tiene en promedio un 38.5 % (ver columna PG_H de la Tabla 5.27) y los expertos tienen un promedio de 91 % (ver columna PE_H de la Tabla 5.28). En esta categoría de medidas al igual que las medidas basadas en contenido de información (dado que las medidas híbridas toman en cuenta el contenido de información para su cálculo) no se tiene resultados favorables ya que se tiene una diferencia considerable de 52.5 %.

Los resultados de promedios generales del algoritmo (PG) en las medidas basadas en contenido de información y las medidas híbridas obtienen valores bajos en comparación a los promedios de los expertos (PE) se debe a que los resultados de similitud semántica obtenidos por cada medida por cada par de conceptos que integran la relación taxonómica influyen en el resultado del cálculo de PG cuando se

comparan con el umbral promedio ($Umb_{promedio}$) (ver Sección 4.1.5)

En general, con base en los resultados de las tablas 5.27 y 5.28 se observa que el Algoritmo tiene mejor eficacia usando las medidas basadas en estructura y las medidas basadas en características al tener a lo mínimo un 80 % de exactitud en promedio con las tres ontologías. Por otro lado, con las medidas basadas en contenido de información y las medidas híbridas no se tiene resultados favorables en el Algoritmo 1 ya que sus resultados van por debajo del 50 % en promedio. Al final de cada una de las tablas se tiene el promedio de $PG_{promedio}$ y $PE_{promedio}$, donde se tiene 55.4 % y 84.2 % respectivamente, este es un promedio tomando en cuenta todas las categorías de medidas y las tres ontologías, mediante estos valores se tiene una diferencia del Algoritmo 1 con los expertos de 28.8 %, siendo así que el algoritmo tiene un margen de error de 28 % en la evaluación de relación de relaciones taxonómicas en comparación a los tres expertos.

5.3. Evaluación de ontologías de dominio de biomedicina

En esta evaluación se toman los resultados de la investigación hecha en [26], donde se evaluaron ontologías del dominio de biomedicina para crear un sistema de recomendación mejorado para NCBO BioPortal[2]. El objetivo de este sistema de recomendación es sugerir la mejor ontología para los datos de entrada que se le proporciona. El sistema original lo que hace es aceptar datos de entrada, ya sea una lista de palabras clave o un corpus textual. Al tener estos datos de entrada, se invoca al anotador de NCBO, que anota los datos textuales con clases de las ontologías. Posteriormente, el sistema califica todas las ontologías en función al número y a la relevancia de las anotaciones encontradas. Por último, muestra las ontologías ordenadas siendo la primera la que el sistema considera más apropiada para los datos de entrada. El sistema de recomendación v2.0 propuesto en [26], agrega una nueva característica que es la evauación de las ontologías, donde se consideran los siguientes criterios: cobertura, aceptación, detalle y especialización. En la evaluación del sistema de recomendación v2.0, se hizo un experimento donde se evalúa qué tan bien el

[2]https://bioportal.bioontology.org/

sistema prioriza la recomendación de ontologías pequeñas que brindan una cobertura apropiada para los datos de entrada. Se crearon 15 datos de entrada que contienen palabras clave de dominios muy específicos y se ejecutaron las dos versiones del sistema de recomendación con esos datos de entrada. En la Tabla 5.29 se muestra el dominio de las ontologías usadas, y la ontología recomendada por cada sistema de recomendación.

Tabla 5.29: Ontologías sugeridas por cada uno de los sistemas de recomendación

Dominio	Resultado v1.0	Resultado v2.0
Investigaciones biomédicas	NCIT	OBI
Signos y Síntomas	NCIT	SYMP
Tipos de celulas	SYN	CL

Estas mismas ontologías fueron usadas para hacer la evaluación de sus relaciones taxonómicas y para verificar con base en los resultados obtenidos de la Tabla 5.29 si los resultados del algoritmo propuesto coinciden con el sistema de recomendación versión 2.0. A continuación se muestran los resultados de la aplicación de nuestro sistema con las ontologías de la Tabla 5.29.

5.3.1. Resultados de las medidas basadas en estructura

A continuación, se presentan los resultados obtenidos de la evaluación de relaciones taxonómicas de las ontologías en el dominio de biomedicina usando las medidas de similitud basadas en estructura. Los umbrales para cada medida (Umb_{sim_i}) y los umbrales promedio ($Umb_{promedio}$) se presentan en la Tabla 5.30.

Tabla 5.30: Umbrales de las medidas basadas en estructura en las ontologías de biomedicina

Ontología	$Umb_{sim_{Path}}$	$Umb_{sim_{WUP}}$	$Umb_{sim_{Li}}$	$Umb_{promedio}$
SYN	0.5	0.912	0.810	0.793
CL	0.5	0.798	0.781	0.693
OBI	0.5	0.890	0.802	0.731
SYMP	0.5	0.892	0.810	0.731
NCIT	0.5	0.899	0.808	0.736

Las exactitudes obtenidas con cada medida (Ex_{sim_i}), la exactitud promedio general del algoritmo (PG) al evaluar las relaciones taxonómicas se presentan en la Tabla 5.31.

Tabla 5.31: Exactitudes de las medidas basadas en estructura para las ontologías de biomedicina

Ontología	$Ex_{sim_{Path}}$	$Ex_{sim_{WUP}}$	$Ex_{sim_{Li}}$	PG
SYN	1.0	0.453	0.766	0.766
CL	1.0	0.816	0.880	0.816
OBI	1.0	0.702	0.937	0.898
SYMP	1.0	0.413	0.810	0.413
NCIT	1.0	0.750	0.781	0.750

En base sólo a la evaluación de las relaciones taxonómicas con las medidas basadas en estructura con las exactitudes obtenidas, y en comparación a los resultados del sistema de recomendación v2.0 del trabajo de Martínez-Romero et al. [26] (ver Tabla 5.5), se obtienen los resultados del algoritmo propuesto en la tabla 5.32 se muestran los resultados en comparación con el sistema de recomendación v2.0.

Tabla 5.32: Dominio de entrada y resultados de los sistemas

Dominio	Resultado v2.0	Resultado algoritmo propuesto
Investigaciones biomédicas	OBI	OBI
Signos y Síntomas	SYMP	NCIT *
Tipos de celulas	CL	CL

Con base en los resultados y a los dominios de las ontologías (ver Sección 5.1.2), se tiene que para el dominio de Investigaciones biomédicas la ontología que obtuvo un promedio general (PG) mayor es OBI con un 89.8 % en comparación con NCIT que obtuvo un 75 %. En el dominio de Signos y Síntomas la ontología que obtuvo un PG mayor es NCIT con un 75 % en comparación a la ontología SYMP que obtuvo un 41.3 %. En el dominio de Tipos de células, la ontología que obtuvo un PG mayor es CL con un 81.6 % en comparación a SYN que obtuvo un 76.6 %.

El sistema de recomendación v2.0 y el algoritmo propuesto en esta tesis coinciden en dos recomendaciones, sólamente con el dominio de Signos y Síntomas no coinciden, esto se debe a que la ontología NCIT tiene mayor número de conceptos y de relaciones

taxonómicas en comparación a SYMP que dentro de las ontologías consideradas para este conjunto de datos tiene menor número de conceptos y relaciones taxonómicas (ver Tabla 5.4).

5.3.2. Resultados de las medidas basadas en características

A continuación, se presentan los resultados obtenidos al evaluar las relaciones taxonómicas de las ontologías en el dominio de biomedicina usando las medidas basadas en características. Los umbrales de cada medida (Umb_{sim_i}) y los umbrales promedio ($Umb_{promedio}$) se presentan en la Tabla 5.33.

Tabla 5.33: Umbrales de cada medida basadas en características en las ontologías de biomedicina

Ontología	$Umb_{sim_{CMatch}}$	$Umb_{sim_{RE}}$	$Umb_{sim_{SAN}}$	$Umb_{promedio}$
SYN	0.806	0.952	0.254	0.669
CL	0.752	0.870	0.301	0.641
OBI	0.801	0.932	0.254	0.662
SYMP	0.759	0.940	0.311	0.669
NCIT	0.790	0.940	0.270	0.667

Las exactitudes que se obtuvieron de cada medida (Ex_{sim_i}) y la exactitud promedio general del algoritmo (PG) para cada ontología del dominio de biomedicina se presentan en la Tabla 5.34.

Tabla 5.34: Exactitudes de las medidas basadas en características para las ontologías de biomedicina

Ontología	$Ex_{sim_{CMatch}}$	$Ex_{sim_{RE}}$	$Ex_{sim_{SAN}}$	PG
SYN	0.438	0.773	0.562	0.773
CL	0.673	0.842	0.327	0.842
OBI	0.647	0.898	0.353	0.937
SYMP	0.404	0.404	0.596	0.851
NCIT	0.758	0.758	0.242	0.844

Con base en los resultados, se tiene que en el dominio de Investigaciones biomédicas la ontología que obtuvo un promedio general (PG) mayor es OBI con un 93.7 %

en comparación con NCIT que obtuvo un 84.4 %. En el dominio de Signos y Síntomas la ontología que obtuvo un PG mayor es SYMP con un 85.1 % en comparación a la ontología NCIT que obtuvo un 84.4 %. En el dominio de Tipos de células, la ontología que obtuvo un PG mayor es CL con un 84.2 % en comparación a SYN que obtuvo un 77.3 %.

La comparación de los resultados del sistema de recomendación v2.0 con el algoritmo propuesto se presenta en la Tabla 5.35

Tabla 5.35: Dominio de entrada y resultados de los sistemas

Dominio	Resultado v2.0	Resultado sistema propuesto
Investigaciones biomédicas	OBI	OBI
Signos y Síntomas	SYMP	NCIT *
Tipos de celulas	CL	CL

Con base en los resultados obtenidos y a los dominios de las ontologías (ver Sección 5.1.2), se tiene que con las medidas basadas en características, sólo en el dominio de signos y Síntomas hay una diferencia en los resultados, donde el Algoritmo 1 propone a NCIT como la ontología mejor estructurada en comparación a SYMP en base a la evaluación de sus relaciones taxonómicas obteniendo mejor valor de exactitud (PG) como se ve en la Tabla 5.34, esto se debe a que la diferencia en el número de relaciones taxonómicas entre las ontologías NCIT y SYMP es considerable como se muestra en la Tabla 5.4.

5.3.3. Resultados de las medidas basadas en contenido de información

En esta sección se presentan los resultados obtenidos de la evaluación de las relaciones taxonómicas utilizando las medidas de similitud basadas en contenido de información. En la Tabla 5.36 se muestran los umbrales para cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$).

Posteriormente, se muestra en la Tabla 5.37 las exactitudes que se obtuvieron para cada medida de similitud con cada ontología (Ex_{sim_i}) y la exactitud general promedio que obtiene el algoritmo propuesto (PG) al evaluar cada ontología.

Tabla 5.36: Umbrales de las medidas basadas en contenido de información en las ontologías de biomedicina

Ontología	$Umb_{sim_{Lin}}$	$Umb_{sim_{Maz}}$	$Umb_{sim_{JC}}$	$Umb_{sim_{Res}}$	$Umb_{promedio}$
SYN	0.841	0.737	0.899	0.567	0.761
CL	0.893	0.819	0.926	0.678	0.829
OBI	0.864	0.770	0.905	0.655	0.798
SYMP	0.823	0.709	0.881	0.600	0.753
NCIT	0.855	0.756	0.902	0.621	0.784

Tabla 5.37: Exactitudes de las medidas basadas en contenido de información para las ontologías de biomedicina

Ontología	$Ex_{sim_{Lin}}$	$Ex_{sim_{Maz}}$	$Ex_{sim_{JC}}$	$Ex_{sim_{Res}}$	PG
SYN	0.564	0.536	0.521	0.517	0.540
CL	0.656	0.634	0.611	0.598	0.628
OBI	0.603	0.570	0.571	0.565	0.583
SYMP	0.572	0.542	0.529	0.542	0.537
NCIT	0.567	0.540	0.528	0.518	0.766

Con base en los resultados, se tiene que en el dominio de Investigaciones biomédicas la ontología que obtuvo un promedio general (PG) mayor es NCIT con un 76.6 % en comparación con OBI que obtuvo un 58.3 %. En el dominio de Signos y Síntomas la ontología que obtuvo un PG mayor es NCIT con un 76.6 % en comparación a la ontología SYMP que obtuvo un 53.7 %. En el dominio de Tipos de células, la ontología que obtuvo un PG mayor es CL con un 62.8 % en comparación a SYN que obtuvo un 54 %.

Finalmente, en la Tabla 5.38 se presentan las ontologías que sugiere el sistema de recomendación v2.0 y las ontologías que tienen mayor exactitud en nuestro sistema en cada dominio.

Tabla 5.38: Dominio de entrada y resultados de los sistemas

Dominio	Resultado v2.0	Resultado sistema propuesto
Investigaciones biomédicas	OBI	NCIT *
Signos y Síntomas	SYMP	NCIT *
Tipos de celulas	CL	CL

Con base en los resultados y a los dominios de las ontologías (ver Sección 5.1.2) se tiene que con las medidas basadas en contenido de información sólo se tiene una coincidencia en el dominio de tipos de células. En los dominios de investigaciones biomédicas y signos y Síntomas la ontología que el Algoritmo 1 propone como mejor ontología para ese dominio en base a la evaluación de las relaciones taxonómicas es NCIT, esto se debe a que esta ontología tiene mayor información como número de conceptos, relaciones taxonómicas y profundidad que son aspectos considerados para el cálculo del contenido de información en comparación con las ontologías OBI y SYMP (ver Tabla 5.4).

5.3.4. Resultados de las medidas híbridas

En esta sección se presentan los resultados obtenidos de la evaluación de las relaciones taxonómicas de las ontologías del dominio de biomedicina utilizando las medidas híbridas. En la Tabla 5.50 se muestran los umbrales para cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$).

Tabla 5.39: Umbrales de las medidas híbridas en las ontologías de biomedicina

Ontología	$Umb_{sim_{Zhou}}$	$Umb_{sim_{HSS3}}$	$Umb_{sim_{HSS4}}$	$Umb_{promedio}$
SYN	0.083	0.040	0.062	0.062
CL	0.044	0.116	0.178	0.113
OBI	0.109	0.080	0.124	0.105
SYMP	0.143	0.105	0.164	0.137
NCIT	0.106	0.073	0.112	0.097

Posteriormente, en la Tabla 5.51 se muestran las exactitudes que se obtuvieron para cada medida de similitud con cada ontología (Ex_{sim_i}) y la exactitud promedio que obtiene el algoritmo propuesto al evaluar cada ontología (PG).

Con base en los resultados y en el dominio considerado para cada ontología (ver Sección 5.1.2), se tiene que en el dominio de Investigaciones biomédicas la ontología que obtuvo un promedio general (PG) mayor es OBI con un 47 % en comparación con NCIT que obtuvo un 46.6 %. En el dominio de Signos y Síntomas la ontología que obtuvo un PG mayor es SYMP con un 54.7 % en comparación a la ontología NCIT que obtuvo un 46.6 %. En el dominio de Tipos de células, la ontología que

Tabla 5.40: Exactitudes de las medidas híbridas para las ontologías de biomedicina

Ontología	$Ex_{sim_{Zhou}}$	$Ex_{sim_{HSS3}}$	$Ex_{sim_{HSS4}}$	PG
SYN	0.521	0.467	0.467	0.458
CL	0.611	0.492	0.490	0.511
OBI	0.571	0.449	0.448	0.470
SYMP	0.529	0.533	0.553	0.547
NCIT	0.528	0.438	0.455	0.466

obtuvo un PG mayor es CL con un 51.1 % en comparación a SYN que obtuvo un 45.8 %.

Finalmente, en la Tabla 5.41 se presentan las ontologías que sugiere el sistema de recomendación v2.0 y las ontologías que tienen mayor exactitud en el algoritmo propuesto en cada dominio.

Tabla 5.41: Dominio de entrada y resultados de los sistemas

Dominio	Resultado v2.0	Resultado sistema propuesto
Investigaciones biomédicas	OBI	OBI
Signos y síntomas	SYMP	SYMP
Tipos de celulas	CL	SYN

Con las medidas híbridas tanto las ontologías sugeridas por el sistema de recomendación v2.0 y los resultados de la evaluación hecha por el algoritmo propuesto coinciden, a pesar de obtener valores muy bajos en los resultados de similitud semántica para esta categoría de medidas, el algoritmo concuerda con el sistema de recomendación v2.0.

5.3.5. Resultados en todas las categorías de las medidas

En esta sección se presenta una recopilación de las exactitudes promedio obtenidas por cada categoría de medidas de similitud usadas en este trabajo por cada ontología para la evaluación de las relaciones taxonómicas con el objetivo de hacer una discusión de los resultados y así observar el comportamiento de cada categoría de medida con las ontologías y de los resultados obtenidos al realizar la comparación con el sistema de recomendación v2.0.

En la tabla 5.42 se muestran las exactitudes promedio para cada categoría de medidas en todas las ontologías del dominio de biomedicina usadas en este trabajo. Al final de la tabla se muestra un promedio por cada categoría de medida y tambíen el promedio de $PG_{promedio}$ que toma en cuenta todas las categorías de medidas y todas las ontologías de biomedicina. Se puede observar que sólo las medidas híbridas obtienen un promedio por debajo del 50 %.

Tabla 5.42: Exactitud para todas las categorías

Ontología	PG_{Str}	PG_{Ft}	PG_{CI}	PG_{H}	$PG_{promedio}$
SYN	0.766	0.773	0.540	0.458	0.634
CL	0.816	0.842	0.628	0.511	0.699
OBI	0.898	0.937	0.583	0.470	0.722
SYMP	0.413	0.851	0.537	0.547	0.587
NCIT	0.750	0.844	0.766	0.466	0.707
Promedio	0.729	0.849	0.611	0.490	0.670

Con los datos de la tabla 5.42, se tiene los siguientes resultados por cada dominio:

- Investigaciones biomédicas: NCIT tiene un promedio ($PG_{promedio}$) de 70.7 % y OBI tiene un 72.2 %, teniendo la ontología OBI un tiene una mejor evaluación en este dominio con todas las categorías de medidas.

- Signos y Síntomas: SYMP tiene un $PG_{promedio}$ de 58.7 % y NCIT cuenta con un promedio de 70.7 %, en este dominio la ontología NCIT obtiene una mejor evaluación de sus relaciones con todas las categorías de medidas.

- Tipos de células: SYN obtiene un $PG_{promedio}$ de 63.4 % y la ontología CL obtiene un promedio de 69.9 %, siendo la ontología CL quien tiene mejor evaluación en sus relciones en comparación de SYN en este dominio con todas las medidas.

En la tabla 5.43 se muestran los resultados en comparación del Algoritmo 1 y los sistemas de recomendación de NCBO versión 1 y 2 por cada dominio. Se puede observar que el sistema de recomendación NCBO de Martínez-Romero et al. [26] coincide en los dominios Investigaciones biomédicas y Tipos de células con el Algoritmo 1 que evalúa solamente las relacions taxonómicas de las ontologías. En el dominio de Signos y Síntomas difieren el Algoritmo 1 y el sistema de recomedación v2.0 esto se

debe a que la ontología SYMP es de un tamaño mucho menor a NCIT, contando con poco más de 900 conceptos y NCIT tiene más de 100,000 conceptos (ver Tabla 5.4), por lo que para la ontología NCIT se tiene más información requerida para el cálculo de contenido de información.

Tabla 5.43: Dominio de entrada y resultados de los sistemas

Dominio	v2.0	Algoritmo 1
Investigaciones biomédicas	OBI	OBI
Signos y Síntomas	SYMP	NCIT
Tipos de células	CL	CL

5.3.6. Discusión

En esta sección se discutirán los resultados obtenidos al evaluar las relaciones taxonómicas en las ontologías del dominio de biomedicina, discutiendo el comportamiento de las medidas de similitud dentro de su categoría.

En la tabla 5.42, tomando en cuenta los promedios al final de la tabla por cada categoría de las medidas de similitud se tiene los siguientes resultados:

- Medidas basadas en estructura: tiene un promedio de 72.9 %
- Medidas basadas en características: tiene un promedio de 84.9 %
- Medidas basadas en contenido de información: tiene un promedio de 61.1 %
- Medidas basadas híbridas: tiene un promedio de 49 %

Con las ontologías del dominio de biomedicina siendo de tamaño mayor a las ontologías del primer conjunto de datos (ver Tabla 5.27) teniendo en total desde 900 hasta más de 100,000 conceptos, se tiene que las medidas basadas en estructura, las basadas en características y las basadas en contenido de información obtienen un promedio por arriba del 60 % mientras que en las medidas híbridas se tiene un promedio por debajo del 49 %. Con ontologías de este tamaño las medidas basadas en características como las basadas en estructura son las que obtienen resultados cercanos con una diferencia de 12 % y las dos categorías están por arriba de un 70 %, por lo que estas medidas obtienen resultados favorables con el Algoritmo 1.

5.4. Resultados con el conjunto de datos en el dominio de sensores

En esta sección, se aplican las medidas de similitud semántica basadas en conocimiento presentadas en la Sección 3.5.3 para evaluar las relaciones taxonómicas de las ontologías en el dominio de sensores.

5.4.1. Resultados de las medidas basadas en estructura

En esta subsección se presentan los resultados obtenidos con las medidas basadas en estructura usando la biblioteca *SML* y que aplica las ecuaciones 3.1, 3.4 y 3.6 para las ontologías en el dominio de sensores. En la tabla 5.44, se muestran el umbral promedio ($Umb_{promedio}$), y el umbral calculado para cada medida de similitud (Umb_{sim_i}).

Tabla 5.44: Umbrales de las medidas basadas en estructura en las ontologías de sensores

Ontología	$Umb_{sim_{Path}}$	$Umb_{sim_{WUP}}$	$Umb_{sim_{Li}}$	$Umb_{promedio}$
Intelligent Environment	0.500	0.775	0.632	0.636
Onto Sensor	0.500	0.794	0.668	0.654
Sensor Ontology	0.500	0.442	0.355	0.433
Device Ontology	0.500	0.757	0.601	0.619

En la tabla 5.45, se presentan las exactitudes obtenidas para cada medida de similitud por cada ontología(Ex_{sim_i}), usando la ecuación 4.3. También en esta tabla se presenta la exactitud promedio general obtenido de cada ontología (PG).

Tabla 5.45: Exactitudes de las medidas basadas en estructura para las ontologías del dominio de sensores

Ontología	$Ex_{sim_{Path}}$	$Ex_{sim_{WUP}}$	$Ex_{sim_{Li}}$	PG
Intelligent Environment	1.0	0.852	0.852	0.852
Onto Sensor	1.0	0.755	0.755	0.755
Sensor Ontology	1.0	0.578	0.578	0.578
Device Ontology	1.0	0.600	0.600	0.600

Como se observa en la tabla 5.45, la ontología con el mayor valor en la exactitud promedio general (PG) del sistema es Intelligent Environment obteniendo un 85.2 % para esta categoría de medidas, lo que quiere decir es que con la ontología Intelligent Environment se obtiene un mejor promedio de exactitud usando las tres medidas basadas en estructura en su evaluación de relaciones taxonómicas.

5.4.2. Resultados de las medidas basadas en características

En esta subsección se muestran los resultados obtenidos para las medidas basadas en características que toman en cuenta los ancestros de los conceptos u y v comparados y que se obtienen usando funciones de la biblioteca *SML* aplicando las ecuaciones 3.8, 3.10 y 3.12.

En la tabla 5.46 se presentan los umbrales obtenidos para cada medida de similitud por cada ontología (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$).

Tabla 5.46: Umbrales de las medidas basadas en características en las ontologías de sensores

Ontología	$Umb_{sim_{CMatch}}$	$Umb_{sim_{RE}}$	$Umb_{sim_{SAN}}$	$Umb_{promedio}$
Intelligent Environment	0.752	0.858	0.318	0.643
Onto Sensor	0.786	0.875	0.273	0.645
Sensor Ontology	0.634	0.769	0.444	0.616
Device Ontology	0.726	0.840	0.348	0.638

En la tabla 5.47 se presentan las exactitudes obtenidas para cada medida (Ex_{sim_i}) por cada ontología usando la ecuación 4.3, además de la exactitud promedio general por cada ontología (PG).

Tabla 5.47: Exactitudes de las medidas basadas en características para las ontologías del dominio de sensores

Ontología	$Ex_{sim_{CMatch}}$	$Ex_{sim_{RE}}$	$Ex_{sim_{SAN}}$	PG
Intelligent Environment	0.296	0.296	0.704	0.852
Onto Sensor	0.566	0.566	0.434	0.570
Sensor Ontology	0.578	0.578	0.422	0.578
Device Ontology	0.600	0.600	0.400	0.600

Para esta categoría de medidas, la ontología con mayor valor de promedio general es nuevamente Intelligent Environment con un 85.2 % indicando que para las medidas basadas en estructura y características, es Intelligent Environment la ontología que obtiene el mejor promedio de exactitud al evaluar sus relaciones taxonómicas.

5.4.3. Resultados de las medidas basadas en contenido de información

En la categoría de medidas basadas en contenido de información, donde se toman en cuenta la información de cada concepto (profundidad del concepto, profundidad de la taxonomía, etc.) se presenta en esta subsección los resultados obtenidos usando la biblioteca *SML* aplicando las ecuaciones 3.15, 3.17, 3.20 y 3.22.

En la tabla 5.48 se muestran los umbrales obtenidos para cada medida de similitud por ontología (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$).

Tabla 5.48: Umbral de las medidas basadas en contenido de información en las ontologías de sensores

Ontología	$Umb_{sim_{Lin}}$	$Umb_{sim_{Maz}}$	$Umb_{sim_{JC}}$	$Umb_{sim_{Res}}$	$Umb_{promedio}$
Intelligent Environment	0.864	0.764	0.895	0.698	0.806
Onto Sensor	0.842	0.750	0.908	0.592	0.773
Sensor Ontology	0.558	0.454	0.803	0.376	0.548
Device Ontology	0.846	0.735	0.889	0.614	0.771

En la tabla 5.49 se presentan la exactitud promedio general (PG) y las exactitudes que se obtienen para cada medida de similitud (Ex_{sim_i}) por ontología usando la ecuación 4.3.

Tabla 5.49: Exactitud de las medidas basadas en contenido de información en las ontologías del dominio de sensores

Ontología	$Ex_{sim_{Lin}}$	$Ex_{sim_{Maz}}$	$Ex_{sim_{JC}}$	$Ex_{sim_{Res}}$	PG
Intelligent Environment	0.593	0.444	0.481	0.444	0.444
Onto Sensor	0.642	0.608	0.585	0.532	0.581
Sensor Ontology	0.578	0.578	0.578	0.578	0.780
Device Ontology	0.300	0.300	0.350	0.450	0.475

Observando la tabla 5.49 tenemos que la ontología con mejor valor en el promedio

general (PG) es Sensor Ontology con un 78 %, teniendo cambios en comparación con los resultados en las medidas basadas en estructura y características porque la ontología Intelligent Environment obtuvo los mejores valores de promedio general (PG) en los resultados de las tablas 5.45 y 5.47 arriba del 85 %, en cambio, con los resultados de la tabla 5.49 se obtienen valores por debajo del 50 %. Esto se debe a que las medidas basadas en contenido de información toman en cuenta aspectos como la profundidad tanto de la taxonomía como de los conceptos comparados, el número total de conceptos, y los descendientes de cada concepto comparado y la ontología Intelligent Environment al ser una ontología muy pequeña con un total de conceptos menor a 100 (ver Tabla 5.6) en comparación a las demás ontologías.

5.4.4. Resultados de las medidas híbridas

En la categoría de medidas híbridas donde es una combinación de medidas basadas en estructura y basadas en contenido de información, se presentan los resultados obtenidos al aplicar el sistema usando funciones de la biblioteca *SML* que usa las ecuaciones 3.24, 3.26 y 3.28.

En la tabla 5.50 se muestran los umbrales que obtiene el sistema con base en los resultados de similitud de cada medida (Umb_{sim_i}) y el umbral promedio ($Umb_{promedio}$).

Tabla 5.50: Umbrales para las medidas híbridas

Ontología	$Umb_{sim_{Zhou}}$	$Umb_{sim_{HSS3}}$	$Umb_{sim_{HSS4}}$	$Umb_{promedio}$
Intelligent Environment	0.126	0.415	0.615	0.385
Onto Sensor	0.065	0.190	0.292	0.182
Sensor Ontology	0.059	0.143	0.221	0.141
Device Ontology	0.102	0.243	0.368	0.238

En la tabla 5.51 se presentan la exactitud promedio general (PG) y las exactitudes que se obtienen para cada medida de similitud (Ex_{sim_i}) por ontología usando la ecuación 4.3.

Observando la tabla 5.51, la ontología con mayor valor en el promedio general (PG) es Device Ontology con un 60 %. En esta categoría de medidas híbridas se tienen resultados diferentes a los vistos en las tablas 5.45, 5.47 pero comparte valores cercanos con 5.49, esto se debe a que las medidas híbridas miden aspectos tanto de

Tabla 5.51: Exactitudes de las medidas híbridas para las ontologías del dominio de sensores

Ontología	$Ex_{sim_{Zhou}}$	$Ex_{sim_{HSS3}}$	$Ex_{sim_{HSS4}}$	PG
Intelligent Environment	0.481	0.296	0.296	0.296
Onto Sensor	0.585	0.370	0.370	0.381
Sensor Ontology	0.578	0.359	0.359	0.375
Device Ontology	0.350	0.600	0.600	0.600

las medidas basadas en estructura y contenido de información, pero tiene más peso el contenido de información en estas medidas.

5.4.5. Resultados en todas las categorías de medidas

En la tabla 5.52, se muestran los promedios generales obtenidos en cada categoría (medidas basadas en estructura (PG_{Str}), medidas basadas en características (PG_{Ft}), medidas basadas en contenido de información (PG_{CI}) y las medidas híbridas (PG_H)), tambien se muestra el promedio por cada ontología de los promedios generales (PG) en todas las categorías $PG_{promedio}$.

Tabla 5.52: Exactitud para todas las categorías

Ontología	PG_{Str}	PG_{Ft}	PG_{CI}	PG_H	$PG_{promedio}$
Intelligent Environment	0.852	0.852	0.444	0.296	0.611
Onto Sensor	0.755	0.570	0.581	0.381	0.571
Sensor Ontology	0.578	0.578	0.780	0.375	0.577
Device Ontology	0.600	0.600	0.475	0.600	0.568
Promedio	0.696	0.650	0.570	0.413	0.582

Como se observa en la tabla 5.52, Intelligent Environment en promedio tiene los mejores resultados en la evaluación de las relaciones taxonómicas con un 61 % de $PG_{promedio}$, aunque con las medidas basadas en contenido de información e híbridas se tienen los PG más bajos con un 44.4 % y 29.6 % respectivamente.

5.4.6. Discusión

Mediante los resultados presentados en la tabla 5.52, tanto las medidas basadas en estructura como las basadas en características se obtienen resultados de PG iguales en cada ontología. Las medidas basadas en contenido de información tienen resultados favorables con ontologías con más profundidad (más de 10 niveles) y conceptos en ellas (con 200 clases o más) como con la ontología Onto Sensor con un PG_{CI} del 78 %. Con las medidas híbridas se obtienen valores muy bajos por debajo del 50 % (a excepción de la ontología Device Ontology con un 60 %), por lo que las medidas híbridas usadas en esta tesis no son favorables para la evaluación de relaciones taxonómicas.

Las ontologías en el dominio de sensores usadas en este experimento son en su mayoría pequeñas con menos de 100 clases, a excepción de Onto Sensor con más de 200 clases, como se ve en la tabla 5.6. Con los resultados experimentales, podemos indicar que cuando las ontologías de sensores son del tamaño similar a ontologías como Intelligent Environment en su evaluación de las relaciones taxonómicas es recomendable usar las medidas basadas en estructura y las basadas en características. Cuando las ontologías son más grandes como OntoSensor es recomendable usar las medidas basadas en contenido de información al contar de mayor información para cada concepto.

5.5. Discusión general

La evaluación de ontologías es un proceso importante para verificar la calidad de las ontologías. Los desarrolladores de ontologías deben tener herramientas para la evaluación que sean fáciles de aplicar, es decir, que no necesite en su totalidad del desarrollador para realizar el proceso manualmente y que se necesite del menor tiempo posible para aumentar la efectividad. En esta tesis se usan medidas de similitud para la evaluación de relaciones taxonómicas en ontologías de dominio. El algoritmo usado es simple y fácil de usar, ya que solo se necesita la ontología en formato OWL y el sistema se encarga de realizar el preprocesamiento y evaluación de la ontología y al final muestra los resultados obtenidos. La finalidad de usar varias categorías de medidas de similitud es ayudar al desarrollador a elegir la mejor categoría que se

adapte a la ontología a evaluar. Con base en los resultados experimentales en los tres conjuntos de datos se tiene que en el Algoritmo 1 propuesto las medidas que son recomendables usar son las medidas basadas en estructura y las medidas basadas en características pero en ontologías con al menos 200 conceptos (como en el primer ver Tabla 5.2 y tercer conjunto de datos, ver Tabla 5.6) . Usando las medidas basadas en contenido de información, a pesar de comportarse mejor en ontologías con mayor número de conceptos (arriba de 900, ver Tabla 5.4) no se comporta al nivel de las medidas basadas en estructura y características (ver Tabla 5.42) y en los otros conjuntos de datos (ver Tabla 5.42 y 5.52) obtiene promedios muy bajos cerca de un 50 %. Con las medidas híbridas en todos los conjuntos de datos se obtienen promedios por debajo del 50 % por lo que estas medidas no son recomendables para la evaluación de las relaciones taxonómicas.

Conclusiones

En este trabajo de investigación se evalúan los conceptos de las relaciones taxonómicas con medidas de similitud semántica basadas en conocimiento para medir la calidad de una ontología de dominio.

Para llevar a cabo esta investigación primero se consultó en la literatura las propuestas que existen de medidas de similitud y se encontró que para las relaciones taxonómicas se utilizan las medidas de similitud basadas en conocimiento, que son aquellas que sólo toman en cuenta la estructura de la taxonomía sin necesitar de un corpus textual. Dentro de las medidas basadas en conocimiento se encuentran diferentes categorías dependiendo de los aspectos que se miden para obtener el grado de similitud de dos conceptos. Estas categorías son: las medidas basadas en estructura, las medidas basadas en características, las medidas basadas en contenido de información y las medidas híbridas. Para este trabajo se implementaron en un algoritmo propuesto algunas de las medidas de cada categoría que son las más reconocidas o usadas en la similitud semántica. Dentro de este algoritmo tambíen se utilizó la medida de exactitud para evaluar las relaciones taxonómicas obteniendo la exactitud de las relaciones correctas de la ontología. Utilizando esta medida también se evalúa el comportamiento de cada medida al obtener su exactitud al ser aplicada para obtener la similitud de los conceptos. Los experimentos realizados utilizando las medidas basadas en estructura se documentaron en el artículo titulado "Medidas de similitud semántica aplicadas a una ontología de dominio"[11] que fue aceptado en el COMIA[3] (Congreso Mexicano de Inteligencia Artificial) 2018. Usando las medidas basadas en características los resultados fueron documentados en el artículo titulado "Medidas de similitud semántica aplicadas a una ontología de dominio"[12] aceptado en el

[3]http://comia2018.smia.mx/

SENIE 2018[4]. Los experimentos realizados utilizando las medidas basadas en contenido de información se documentaron en el artículo "Applying semantic similarity measures based on information content in the evaluation of a domain ontology"[10] aceptado en el MICAI[5] (Mexican International Conference on Artificial Intelligence) 2018 para una publicación de la IEEE. Mediante los resultados experimentales en ontologías de diferentes dominios, se pudo comparar las medidas implementadas y así saber cuáles son más aptas para ciertos dominios dependiendo de sus características. En los resultados se tiene que el agoritmo funciona mejor con ontologías de un tamaño mayor a 200 conceptos, teniendo esto en cuenta, las medidas basadas en estructura y las medidas basadas en características son las que funcionan mejor con el algoritmo propuesto, teniendo en promedio para las medidas basadas en estructura mayor al 73 % y para las medidas basadas en características se tiene en promedio una exactitud arriba del 82 %. Teniendo estos resultados, el algoritmo propuesto para la evaluación de las relaciones taxonómicas puede ser de ayuda para el proceso de evaluación de una ontología utilizando las medidas basadas en estructura y características ya que estas medidas consideran principalmente la distancia y los ancestros en común de los conceptos de las relaciones taxonómicas dado que en las relaciones clase-subclase directas son las medidas que mejor se adecuan considerando sólamente la estructura de la ontología. Como trabajo futuro se puede proponer una medida híbrida que combine aspectos de las medidas basadas en estructura y características. También se puede incorporar otras medidas de similitud semántica aplicadas a las relaciones no taxonómicas por ejemplo las propuestas en los trabajos de Spanakis et al. [44], Siblini y Kosseim [41], Wei y Chang [60] y Pirró y Euzenat [34].

[4]http://senie.azc.uam.mx/
[5]http://www.micai.org/2018/

Bibliografía

[1] Emanuel Aguilar, Darnes Vilariño, Claudia Zepeda, Mireya Tovar, y Beatríz Beltrán. Desarrollo de un modelo para encontrar la similitud semántica multilingüe. *Research in Computing Science*, 99:73–81, 2015. URL http://rcs.cic.ipn.mx/2015_99/Desarrollo%20de%20un%20modelo%20para%20encontrar%20la%20similitud%20semantica%20multilingue.pdf.

[2] Darnes Vilariño Ayala, Mireya Tovar, Beatríz Beltrán, y Saúl León. Un modelo para detectar la similitud semántica entre textos de diferentes longitudes. *Research in Computing Science*, 85:57–64, 2014. URL http://rcs.cic.ipn.mx/2014_85/Un%20modelo%20para%20detectar%20la%20similitud%20semantica%20entre%20textos%20de%20diferentes%20longitudes.pdf.

[3] A. Bandrowski et al. The Ontology for Biomedical Investigations. *PLoS ONE*, 11(4):e0154556, 2016.

[4] Jonathan Bard, Seung Y. Rhee, y Michael Ashburner. An ontology for cell types. *Genome Biol*, 6(2):R21–R21, 2005. ISSN 1474-760X. doi:10.1186/gb-2005-6-2-r21. URL https://www.ncbi.nlm.nih.gov/pubmed/15693950. 15693950[pmid].

[5] Tim Berners-Lee, James Hendler, y Ora Lassila. The semantic web. *Scientific American*, 284(5):34–43, 2001. URL http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21.

[6] Willem Nico Borst. *Construction of Engineering Ontologies for Knowledge Sharing and Reuse.* Tesis Doctoral, University of Twente, Enschede, Netherlands, 1997. Base-search.net (ftunivtwente:oai:doc.utwente.nl:17864).

[7] Hugo Raziel Lasserre Chavez y Mireya Tovar Vidal. Proposal for automatic extraction of taxonomic relations in domain corpus. *Research in Computing Science*, 133:29–39, 2017. URL http://rcs.cic.ipn.mx/2017_133/Proposal%20for%20Automatic%20Extraction%20of%20Taxonomic%20Relations%20in%20Domain%20Corpus.pdf.

[8] C. M. d. O. Rodrigues, F. L. G. de Freitas, y R. R. de Azevedo. An improved logic-based implementation to ontology assessment. En *2017 Brazilian Conference on Intelligent Systems (BRACIS)*, págs. 408–413. 2017. doi: 10.1109/BRACIS.2017.22.

[9] Aldo Gangemi, Carola Catenacci, Massimiliano Ciaramita, Jos Lehmann, y Francesco Bolici. Ontology evaluation and validation an integrated formal model for the quality diagnostic task. 2005.

[10] Aimee Cecilia Hernandez Garcia, Mireya Tovar Vidal, y Jose de Jesus Lavalle Martinez. Applying semantic similarity measures based on information content in the evaluation of a domain ontology. *2018 Seventieth Mexican International Conference on Artificial Intelligence (MICAI) en proceso de publicación*, 2018.

[11] Aimee Cecilia Hernandez Garcia, Mireya Tovar Vidal, y Jose de Jesus Lavalle Martinez. Medidas de similitud semántica aplicadas a una ontología de dominio. *Research in Computing Science*, 147(6):119–131, 2018. ISSN 1870-4069. URL http://www.rcs.cic.ipn.mx/rcs/2018_147_6/.

[12] Aimee Cecilia Hernandez Garcia, Mireya Tovar Vidal, Jose de Jesus Lavalle Martinez, y Ana Patricia Cervantes Marquez. Medidas de similitud basadas en características para la evaluación de relaciones taxonómicas. *Pistas educativas número especial (Difusión del conocimiento)*, 40(140):588–605, 2018. ISSN 2448-847X.

[13] Vladimir Geroimenko. *Dictionary of XML technologies and the semantic web*. Springer-Verlag London, 1 ed$^{\underline{ón}}$., 2004. ISBN 978-0-85729-376-3.

[14] Asunción Gómez-Pérez. *Ontology Evaluation*, págs. 251–273. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-24750-0. doi:10.1007/978-3-540-24750-0_13. URL https://doi.org/10.1007/978-3-540-24750-0_13.

[15] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5):907 – 928, 1995. ISSN 1071-5819. doi:https://doi.org/10.1006/ijhc.1995.1081. URL http://www.sciencedirect.com/science/article/pii/S1071581985710816.

[16] M. Grüninger y M. Fox. Methodology for the Design and Evaluation of Ontologies. En *IJCAI'95, Workshop on Basic Ontological Issues in Knowledge Sharing, April 13, 1995*. 1995. URL http://citeseer.ist.psu.edu/grninger95methodology.html.

[17] N. Guarino. *Formal Ontology in Information Systems: Proceedings of the 1st International Conference June 6-8, 1998, Trento, Italy*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 1st ed$^{\underline{\text{ón}}}$., 1998. ISBN 9051993994.

[18] S. Harispe, S. Ranwez, y S. Janaqi. *Semantic Similarity from Natural Language and Ontology Analysis*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2015. ISBN 9781627054478. URL https://books.google.com.mx/books?id=M6utCQAAQBAJ.

[19] David J. Russomanno, Cartik Kothari, y Omoju A. Thomas. Building a sensor ontology: A practical approach leveraging iso and ogc models. págs. 637–643. 2005.

[20] Jay J. Jiang y David W. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. *CoRR*, cmp-lg/9709008, 1997. URL http://arxiv.org/abs/cmp-lg/9709008.

[21] Y. Li, Z. A. Bandar, y D. Mclean. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions*

on Knowledge and Data Engineering, 15(4):871–882, 2003. ISSN 1041-4347. doi:10.1109/TKDE.2003.1209005.

[22] Elizabeth D. Liddy. Natural language processing. En Miriam Drake, ed., *Encyclopedia of Library and Information Science, 2nd Ed.*, tomo 3. Marcel Decker, Inc., 2001.

[23] Dekang Lin. An information-theoretic definition of similarity. En *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, págs. 296–304. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998. ISBN 1-55860-556-8. URL http://dl.acm.org/citation.cfm?id=645527.657297.

[24] Jonathan M J Derry, Lara M Mangravite, Christine Suver, Matthew D Furia, David Henderson, Xavier Schildwachter, Brian Bot, Jonathan Izant, Solveig Sieberts, Michael R Kellen, y Stephen Friend. Developing predictive molecular maps of human disease through community-based modeling. *Nature genetics*, 44:127–30, 2012. doi:10.1038/ng.1089.

[25] Alexander Maedche y Steffen Staab. Measuring similarity between ontologies. En Asunción Gómez-Pérez y V. Richard Benjamins, eds., *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*, págs. 251–263. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002. ISBN 978-3-540-45810-4.

[26] Marcos Martínez-Romero, Clement Jonquet, Martin J. O'Connor, John Graybeal, Alejandro Pazos, y Mark A. Musen. Ncbo ontology recommender 2.0: an enhanced approach for biomedical ontology recommendation. *Journal of Biomedical Semantics*, 8(1):21, 2017. ISSN 2041-1480. doi:10.1186/s13326-017-0128-y. URL https://doi.org/10.1186/s13326-017-0128-y.

[27] Gaston Mazandu y Nicola Mulder. Information content-based gene ontology semantic similarity approaches: Toward a unified framework theory. 2013:292063, 2013.

[28] Mark A. Musen. The protÉgÉ project: A look back and a look forward. *AI Matters*, 1(4):4–12, 2015. ISSN 2372-3483. doi:10.1145/2757001.2757003. URL http://doi.acm.org/10.1145/2757001.2757003.

[29] Holger Neuhaus y Michael Compton. The semantic sensor network ontology: A generic language to describe sensor assets. 2009.

[30] Emilio J. Nunez, Larry W. Lake, Robert B. Gilbert, Sanjay Srinivasan, Fan Yang, y Mark W. Kroncke. Towards an ontology driven eor decision support system, 2008. Position Paper - W3C Workshop on Semantic Web in Energy Industries; Part I: Oil & Gas.

[31] Leo Obrst, Werner Ceusters, Inderjeet Mani, Steve Ray, y Barry Smith. *The Evaluation of Ontologies*, págs. 139–158. Springer US, Boston, MA, 2007. ISBN 978-0-387-48438-9. doi:10.1007/978-0-387-48438-9_8. URL https://doi.org/10.1007/978-0-387-48438-9_8.

[32] Razan Paul, Tudor Groza, Andreas Zankl, y Jane Hunter. Semantic similarity-driven decision support in the skeletal dysplasia domain. 2012. doi:10.1007/978-3-642-35173-0_11.

[33] Ted Pedersen, Serguei V.S. Pakhomov, Siddharth Patwardhan, y Christopher G. Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288 – 299, 2007. ISSN 1532-0464. doi:https://doi.org/10.1016/j.jbi.2006.06.004. URL http://www.sciencedirect.com/science/article/pii/S1532046406000645.

[34] Giuseppe Pirró y Jérôme Euzenat. A feature and information theoretic framework for semantic similarity and relatedness. En Peter F. Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff Z. Pan, Ian Horrocks, y Birte Glimm, eds., *The Semantic Web – ISWC 2010*, págs. 615–630. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-17746-0.

[35] Joe Raad y Christophe Cruz. A survey on ontology evaluation methods. En *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, IC3K 2015, págs.

179–186. SCITEPRESS - Science and Technology Publications, Lda, Portugal, 2015. ISBN 978-989-758-158-8. doi:10.5220/0005591001790186. URL https://doi.org/10.5220/0005591001790186.

[36] R. Rada, H. Mili, E. Bicknell, y M. Blettner. Development and application of a metric on semantic nets. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(1):17–30, 1989. ISSN 0018-9472. doi:10.1109/21.24528.

[37] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *CoRR*, abs/1105.5444, 2011. URL http://arxiv.org/abs/1105.5444.

[38] M. A. Rodriguez y M. J. Egenhofer. Determining semantic similarity among entity classes from different ontologies. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):442–456, 2003. ISSN 1041-4347. doi:10.1109/TKDE.2003.1185844.

[39] Carlos Rueda, Nancy Galbraith, Robert Morris, Luis Bermudez, Robert Arko, y John Graybeal. The mmi device ontology: Enabling sensor integration. *American Geophysical Union Fall Meeting – Session*, 16:44–8, 2010.

[40] Nuno Seco, Tony Veale, y Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. En *Proceedings of the 16th European Conference on Artificial Intelligence*, ECAI'04, págs. 1089–1090. IOS Press, Amsterdam, The Netherlands, The Netherlands, 2004. ISBN 978-1-58603-452-8. URL http://dl.acm.org/citation.cfm?id=3000001.3000272.

[41] Reda Siblini y Leila Kosseim. Clac: Semantic relatedness of words and phrases. *CoRR*, abs/1708.05801, 2017. URL http://arxiv.org/abs/1708.05801.

[42] Nicholas Sioutos, Sherri de Coronado, Margaret W. Haber, Frank W. Hartel, Wen-Ling Shaiu, y Lawrence W. Wright. Nci thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J. of Biomedical Informatics*, 40(1):30–43, 2007. ISSN 1532-0464. doi:10.1016/j.jbi.2006.02.013. URL http://dx.doi.org/10.1016/j.jbi.2006.02.013.

[43] Thabet Slimani. Description and evaluation of semantic similarity measures approaches. *CoRR*, abs/1310.8059, 2013. URL http://arxiv.org/abs/1310.8059.

[44] G. Spanakis, G. Siolas, y A. Stafylopatis. A hybrid web-based measure for computing semantic relatedness between words. En *2009 21st IEEE International Conference on Tools with Artificial Intelligence*, págs. 441–448. 2009. ISSN 1082-3409. doi:10.1109/ICTAI.2009.64.

[45] Steffen Staab y Rudi Studer. *Handbook on Ontologies*. Springer Publishing Company, Incorporated, 2nd ed$^{\underline{\text{ón}}}$., 2009. ISBN 3540709991, 9783540709992.

[46] Rudi Studer, V.Richard Benjamins, y Dieter Fensel. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering*, 25(1):161 – 197, 1998. ISSN 0169-023X. doi:https://doi.org/10.1016/S0169-023X(97)00056-6. URL http://www.sciencedirect.com/science/article/pii/S0169023X97000566.

[47] David Sánchez, Montserrat Batet, David Isern, y Aida Valls. Ontology-based semantic similarity: A new feature-based approach. *Expert Systems with Applications*, 39(9):7718 – 7728, 2012. ISSN 0957-4174. doi:https://doi.org/10.1016/j.eswa.2012.01.082. URL http://www.sciencedirect.com/science/article/pii/S0957417412000954.

[48] Mireya Tovar. *Evaluación Automática de Ontologías de Dominio Restringido*. Tesis de doctorado en ciencias, Centro Nacional de Investigación y Desarrollo Tecnológico, Departamento de Ciencias Computacionales, 2015.

[49] Mireya Tovar, David Pinto, Azucena Montes, Gabriel González Serna, Darnes Vilariño Ayala, y Beatríz Beltrán. Use of lexico-syntactic patterns for the evaluation of taxonomic relations. En *Pattern Recognition - 6th Mexican Conference, MCPR 2014, Cancun, Mexico, June 25-28, 2014. Proceedings*, págs. 331–340. 2014. doi:10.1007/978-3-319-07491-7_34. URL https://doi.org/10.1007/978-3-319-07491-7_34.

[50] Mireya Tovar, David Pinto, Azucena Montes, y Juan Gabriel González Serna. An approach based in LSA for evaluation of ontological relations on domain corpora. En *Pattern Recognition - 9th Mexican Conference, MCPR 2017, Huatulco, Mexico, June 21-24, 2017, Proceedings*, págs. 225–233. 2017. doi:10.1007/978-3-319-59226-8_22. URL https://doi.org/10.1007/978-3-319-59226-8_22.

[51] Mireya Tovar, David Pinto, Azucena Montes, Juan Gabriel González Serna, y Darnes Vilariño Ayala. Patterns used to identify relations in corpus using formal concept analysis. En *Pattern Recognition - 7th Mexican Conference, MCPR 2015, Mexico City, Mexico, June 24-27, 2015, Proceedings*, págs. 236–245. 2015. doi:10.1007/978-3-319-19264-2_23. URL https://doi.org/10.1007/978-3-319-19264-2_23.

[52] Mireya Tovar, David Pinto, Azucena Montes, y Darnes Vilariño. Determining the degree of semantic similarity using prototype vectors. En Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad, Joaquín Salas Rodríguez, y Gabriella Sanniti di Baja, eds., *Pattern Recognition*, págs. 364–373. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-38989-4.

[53] Amos Tversky. Features of similarity. 84:327–352, 1977.

[54] A. Verma. An abstract framework for ontology evaluation. En *2016 International Conference on Data Science and Engineering (ICDSE)*, págs. 1–6. 2016. doi:10.1109/ICDSE.2016.7823945.

[55] M. Victoria Escandell Vidal. *Apuntes de semántica léxica*. Editorial UNED, 2011.

[56] Mireya Tovar Vidal, David Eduardo Pinto Avendaño, Azucena Montes Rendón, y Juan Gabriel González Serna. A metric for the evaluation of restricted domain ontologies. *Computación y Sistemas*, 22(1), 2018. URL http://www.cys.cic.ipn.mx/ojs/index.php/CyS/article/view/2792.

[57] Mireya Tovar Vidal, David Pinto Avendaño, Azucena Montes Rendón, Juan Gabriel González Serna, y Darnes Vilariño Ayala. Evaluation of ontological

relations in corpora of restricted domain. *Computación y Sistemas*, 19(1), 2015. URL http://cys.cic.ipn.mx/ojs/index.php/CyS/article/view/1954.

[58] Mireya Tovar Vidal, David Pinto, Azucena Montes, Gabriel González Serna, y Darnes Vilariño Ayala. Identification of ontological relations using formal concept analysis. En *Proceedings of the Ninth Latin American Workshop on Logic/Languages, Algorithms and New Methods of Reasoning, Valle de Bravo, Mexico, November 5-7, 2014.*, págs. 1–9. 2014. URL http://ceur-ws.org/Vol-1287/lanmr2014_paper_8.pdf.

[59] Darnes Vilariño, David Pinto, Mireya Tovar, Saúl León, y Esteban Castillo. BUAP: lexical and semantic similarity for cross-lingual textual entailment. En *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, págs. 706–709. 2012. URL http://aclweb.org/anthology/S/S12/S12-1106.pdf.

[60] Tingting Wei y Huiyou Chang. Measuring word semantic relatedness using wordnet-based approach. *Journal of Computers*, 10:252–259, 2015. doi: 10.17706/jcp.10.4.252-259.

[61] F. Wu, Z. Lu, Y. Yan, y J. Gu. Measuring taxonomic relationships in ontologies using lexical semantic relatedness. En *2009 Second International Conference on the Applications of Digital Information and Web Technologies*, págs. 784–789. 2009. doi:10.1109/ICADIWT.2009.5273860.

[62] Zhibiao Wu y Martha Palmer. Verb semantics and lexical selection. *CoRR*, abs/cmp-lg/9406033, 1994. URL http://arxiv.org/abs/cmp-lg/9406033.

[63] Z. Zhou, Y. Wang, y J. Gu. A new model of information content for semantic similarity in wordnet. En *2008 Second International Conference on Future Generation Communication and Networking Symposia*, tomo 3, págs. 85–89. 2008. doi:10.1109/FGCNS.2008.16.

[64] G. Zhu y C. A. Iglesias. Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85, 2017. ISSN 1041-4347. doi:10.1109/TKDE.2016.2610428.

[65] Ganggao Zhu y Carlos Angel Iglesias. Sematch: Semantic entity search from knowledge graph. En *SumPre-HSWI@ESWC*. 2015.

[66] Amal Zouaq, Dragan Gasevic, y Marek Hatala. Linguistic patterns for information extraction in ontocmaps. En *Proceedings of the 3rd International Conference on Ontology Patterns-Volume 929*, págs. 61–72. CEUR-WS. org, 2012.