UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERÍA
DOCTORADO EN CIENCIAS APLICADAS

Marco Ontológico Dinámico Semántico (MODS) para la Web Semántica

Realizado por: Taniana Rodríguez
Tutor: Jose Aguilar

Junio 2015
Mérida-Venezuela

A Dios todo poderoso
A mis padres José y Olga
A mi esposo amado
A mis hijos

Resumen

En este trabajo doctoral se propone un Marco Ontológico Dinámico Semántico (MODS), que permite el análisis y la realización de consultas en lenguaje natural en la Web. La consulta para el MODS, más que una petición de información, es un elemento cargado de información útil para formarse una idea del tipo de usuario, y aproximarse, de manera sucesiva, a una respuesta que cada vez más satisfaga las necesidades del mismo. MODS tiene el desafío de interpretar y formalizar la consulta realizada por el usuario en lenguaje natural, y refinar sus esquemas internos frente a la dinámica de la Web, para tratar futuras consultas (para eso requiere de mecanismos de adaptación)

De manera general, MODS transforma la consulta a un lenguaje ontológico, utilizando sus diferentes componentes: el lexicón, la ontología lingüística, la ontología de tareas y la ontología de dominio. De esta manera, MODS utiliza mecanismos de la semántica ontológica y herramientas del procesamiento del lenguaje natural para el procesamiento de las consultas de los usuarios. Una descripción más detallada de la arquitectura es la siguiente: la ontología de tareas modela las tareas de procesamiento de la consulta en lenguaje natural (análisis léxico–morfológico, análisis sintáctico-semántico, análisis pragmático); la ontología lingüística especifica la gramática del lenguaje español, y cuenta con una extensión de derivaciones coloquiales y con un lexicón para caracterizar al lenguaje español, que a su vez contiene un onomástico; finalmente, la ontología interpretativa modela el conocimiento sobre el contexto especifico del usuario.-

Otro componente clave para la adaptabilidad del MODS a la dinámica de la web y del usuario, es el componente de aprendizaje de ontologías, cuyo fin es permitir que las ontologías evolucionen a la par con la usabilidad del sistema.

Palabras claves: Semántica Ontológica, Aprendizaje de Ontologías, Web Semántica, Procesamiento de Lenguaje Natural.

Agradecimientos

Este trabajo doctoral realizado en la Universidad de Los Andes, Mérida, Venezuela es el resultado de un gran esfuerzo en la cual participaron distintas personas, dándome ánimo, acompañándome en los momentos difíciles y en los momentos de dicha. Este trabajo me ha permitido aprovechar la experiencia de muchas personas que deseo agradecerles.

A Dios, por darme la fortaleza y la sabiduría para terminar con éxito este trabajo.

A mi tutor Jose Aguilar, un especial agradecimiento, por compartir sus conocimientos, su gran sabiduría y paciencia, apoyo y ánimo que me brindó, y por ser un ejemplo a seguir. Estos años de trabajo juntos fueron de gran enriquecimiento en mi persona. Mil gracias y que Dios lo bendiga siempre

A mi esposo Oswaldo Paredes, por su apoyo incondicional. Te amo.

A mis hijos, por su comprensión en mis ausencias, este logro es de ustedes.

A mis amigos Luisa, Niriaska, Code, Eduard, por acompañarme en este proceso

A la Ilustre Universidad de Los Andes.

Al proyecto CDCHT I-1237-10-02-AA, y su proyecto satélite I-1239-10-02-ED, de la Universidad de Los Andes.

Al Programa de Cooperación de Postgrados (PCP) entre Venezuela y Francia, titulado “Tareas de Supervisión y Mantenimiento en Entornos Distribuidos Organizacionales”, contrato Nro. 2010000307

A todos ustedes, mi mayor conocimiento y gratitud.

Índice

Índice de figuras

Índice de Tablas

Capítulo 1: Introducción

En este capítulo se presenta el problema de investigación que ha motivado el desarrollo del presente trabajo de tesis: la interpretación de la consulta realizada por el usuario en lenguaje natural. Este problema cobra importancia significativa dada la necesidad de recuperación de la información que requiere el usuario en el momento de realizar la consulta en el gran repositorio de información que representa la Web actualmente. Se presenta, además, los objetivos planteados en el desarrollo del presente proyecto, el estado de arte de ese problema en particular, y una breve descripción del contenido de los capítulos. Ahora bien, como la tesis tuvo que considerar otros problemas durante su desarrollo (aprendizaje semántico, etc.), en los capítulos siguientes aparecerán los estados de arte vinculado a esos temas cubiertos.

1.1. Introducción

En los últimos años, la Web ha evolucionado desde aquellos primeros usos en el cual los usuarios solo se conectaban a ella para buscar información, hasta la actualidad donde existen comunidades de usuarios (redes sociales, listas de usuarios, etc.) y una gama inmensa de servicios (de búsqueda (Google, Yahoo, etc.), de colaboración, de información (blogs, wikis, etc.), de conexión, etc.) [2], convirtiendo a la actual Web en un gran repositorio de información que contiene documentos de diversos tipos (textos, imágenes, música, etc.). Esta Web es un espacio de intercambio masivo de información, con un formato universal, el lenguaje HTML, que es interpretado por cualquier navegador.

Pese a la gran cantidad de información contenida en la Web, el problema básico en su utilización como repositorio de información es la falta de una estructura semántica que permita interpretar el contenido de la mayoría de la información contenida en ésta. Por lo tanto, el usuario es el que tiene la tarea de interpretar los distintos elementos y secciones que conforman la Web. La causa es porque la información esta expresada en lenguaje natural, orientada a la comprensión por parte del usuario, lo que dificulta computacionalmente los procesos de búsqueda y localización de la información útil para un usuario, a pesar de que actualmente existen buscadores o motores de búsquedas de buena calidad, tales como Yahoo[1] , Google[2] , AltaVista[3] , etc.

En general, para los buscadores es difícil responder a lo que el usuario realmente requiere, porque están basados solo en búsquedas de palabras claves [6]. Así, los buscadores exploran documentos que contengan una o varias de esas palabras, y luego los ordenan según algún criterio propio de cada uno de ellos. Esto conlleva a que muchas veces esas consultas traigan resultados

[1] http://es.yahoo.com/
[2] http://www.google.com/
[3] http://es.altavista.com/

inconsistentes, o documentos que cumplen con el criterio de búsqueda pero no con el interés del usuario. Para los usuarios, encontrar la información deseada o necesaria, es casi como buscar una aguja en un pajar. Adicionalmente, la mayoría de los usuarios actuales quieren utilizar la Web no solamente para encontrar información, sino también para realizar otras tareas, tales como promocionarse, establecer relaciones, entre otras cosas.

En los últimos años se ha venido desarrollando la *Web Semántica*, que es una extensión de la Web actual. La idea general de la Web Semántica es que la información se convierta en conocimiento por medio del uso de una forma estructurada que sea interpretable, tanto por las personas como por las máquinas, logrando con esto que tanto el usuario como las máquinas puedan encontrar la información deseada en la Web [2]. Para ello, en los actuales momentos, la Web se está poblando de ontologías (es una de las formas estructuradas que se está usando), lo cual plantea nuevos retos en los sistemas de recuperación de información.

Otro reto actualmente es que el usuario pueda preguntar a la Web lo que quiere encontrar usando su propio lenguaje, para lo cual es necesario la interpretación de sus consultas en la Web. En particular, para el proceso de interpretación de la consulta en lenguaje natural, es fundamental la integración de lexicones, ontologías, gramáticas, etc., y la incorporación de procesos de análisis del contenido de las consultas.

1.2. Planteamiento del Problema

Partimos de la hipótesis que cualquier ser humano tiene marcos ontológicos, mediante los cuales representa y entiende el mundo que lo rodea. Sus marcos ontológicos no son explícitos, en el sentido de que no se detallan en un documento, ni se organizan de forma jerárquica o matemática. Por ejemplo, todos usamos un marco ontológico en el que “el automóvil” representa un medio de transporte y tiene cuatro ruedas. Ahora bien, podríamos hacernos la siguiente pregunta, ¿Formalizamos este tipo de ontologías? No, sería innecesario, porque los automóviles son tan habituales que todos compartimos la información de lo que son, y nuestro cerebro lo interpreta naturalmente de esa manera. Lo mismo sucede cuando pensamos en el dominio familiar: sabemos que una familia se compone de varios miembros, que un hijo no puede tener más de un padre y una madre biológicos, que los padres tienen o han tenido padres, etc. No necesitamos explicar este conocimiento, pues forma parte de lo que todo el mundo sabe. Sin embargo, cuando se trata de términos poco comunes, o cuando se quiere que estos términos se conviertan en conocimiento procesables por máquinas, se precisa explicitar las ontologías; esto es, desarrollarlas en un documento, o darles una forma que sea inteligible para las máquinas.

Las máquinas carecen de un marco ontológico como con el que nosotros contamos para entender el mundo y comunicarnos; por lo tanto, se necesita definir las ontologías explícitamente. Por ejemplo, si aplicaciones computacionales

intentan comunicarse (traductores automáticos del lenguaje natural, sistemas de extracción de información, sistemas de generación automática de respuestas, bases de conocimiento, etc.), podrían aparecer problemas semánticos que dificultarían o imposibilitarían la comunicación entre ellas. Un marco ontológico permite organizar la información y el conocimiento, de manera que las aplicaciones computacionales puedan interpretar sus significados y aprender los nuevos significados en forma dinámica.

Actualmente se está trabajando bastante en la Web Semántica, con el objetivo de representar el conocimiento adquirido por los seres humanos en forma explícita, para lograr una comunicación eficiente en internet. Para eso, la Web Semántica está utilizando un formato de datos común para expresar el significado de los datos, ontologías para ayudar a las máquinas, y agentes de software para comprender el contenido Web. En específico, para el formato de datos común la Web Semántica está utilizando lenguajes para la descripción de vocabularios semánticos como RDF, RDFS y OWL. Ahora bien, los sistemas que se han desarrollado para la interpretación de las consultas en lenguaje natural no actualizan sus componentes en forma dinámica.

En particular, nuestro trabajo se focaliza en la interpretación de las consultas hechas a la Web en lenguaje natural, para lo cual se propone un Marco Ontológico Semántico Dinámico que usa un conjunto de ontologías como su componente clave para la representación y almacenamiento del conocimiento, y explota su semántica para la interpretación de la consulta realizada en lenguaje natural. Además, dicho Marco Ontológico tiene una serie de mecanismos de aprendizaje que le permiten mantener actualizado sus ontologías (seguir el comportamiento dinámico de su entorno de uso). Así, el Marco Ontológico pretende facilitar la comunicación entre personas, organizaciones y aplicaciones de manera natural, razonar automáticamente, y aprender en forma dinámica.

Esa es la principal motivación de este trabajo de tesis. De esta manera, nuestro sistema debe posibilitar que el usuario realice la consulta en lenguaje natural y sea interpretado por el Marco Ontológico, y además, debe actualizar sus diferentes componentes en forma dinámica, en particular, cuando se realizan las interpretaciones de las consultas, para agregar nueva información semántica que le permita mejor descifrar las necesidades futuras de información de los usuarios.

1.3. Objetivos

Objetivos General

- ✓ Proponer un Marco Ontológico Dinámico Semántico (MODS) para la Web Semántica.

1.3.1. Objetivos Específicos

- ✓ Profundizar en los aspectos teóricos relacionados a los temas de Web Semántica, Aprendizaje de Ontologías y Semántica Ontológica.
- ✓ Describir formalmente las Ontologías utilizadas en el MODS.
- ✓ Caracterizar el proceso de procesamiento de las consultas en lenguaje natural para la Web usando ontologías.
- ✓ Proponer modelos de aprendizaje automático de ontologías.
- ✓ Probar el Marco Ontológico Dinámico Semántico para la Web Semántica en casos de estudio

1.4. Estado del Arte

Entre los trabajos similares que se han realizado a éste, se pueden mencionar los siguientes:

El sistema MESIA [4] facilita la comunicación del usuario con el motor de búsqueda Altavista. El sistema recibe la consulta del usuario escrita en lenguaje natural y la convierte en una consulta booleana. Además, realiza una expansión de la consulta mediante recursos lingüísticos Finalmente, durante el proceso de búsqueda se produce una expansión de los resultados (después de la búsqueda, MESIA incorpora información sobre el dominio, permitiendo la expansión semántica de los resultados. Además, una vez identificado el tema de la consulta, a los resultados obtenidos se le añaden enlaces sobre asuntos relacionados con dicho tema [33]). A continuación se describe brevemente la arquitectura MESIA [33]. El sistema utiliza dos recursos lingüísticos para la expansión de la consulta, que son: ARIES (es un analizador léxico morfológico para el español, el cual está formado por un léxico español de aproximadamente 38500 lemas y 600 morfemas flexivos [15]. Para su integración con MESIA se realizó la traducción de su base léxica a PROLOG), y EWN (es una base de datos semántica en español, que permite obtener sinónimos de cada palabra [33]). Para la integración de ARIES con EWN se usaron patrones sintácticos, que permiten segmentar frases (descomponer la consulta en sintagmas nominales, preposicionales y verbales, con el fin de resolver la ambigüedad del proceso morfológico producido por ARIES, y permitir el análisis sintáctico para extraer los términos relevantes de la consulta). Esos términos luego son expandidos por EWN [5], y el resultado es enviado en paralelo al motor de búsqueda Altavista y a la ontología de MESIA. Dicha ontología, implementada en XML, consiste en un árbol de conceptos relacionados con cada uno de los temas que pueden encontrarse en un dominio dado, lo que permite incorporar información semántica sobre el dominio al proceso de búsqueda. Los nodos son palabras clave de los temas, e incluyen un conjunto de enlaces relacionados con los mismos [45]. Los resultados, tanto del motor de

[4] Modelo Computacional para Extracción Selectiva de Información de textos cortos
[5] EuroWordNet. http://www.illc.uva.nl/EuroWordNet/

búsqueda como de la ontología se ordenan, y se mezclan, para luego ser presentados al usuario.

El sistema QACID[6] se compone de dos núcleos principales: una base de conocimiento generada a partir de pruebas realizadas con usuarios reales; y un módulo de implicación textual. Para crear la base de conocimiento el sistema recoge preguntas de usuarios realizadas sobre un dominio específico. Dichas preguntas son analizadas y agrupadas en función de la información que solicitan. A cada agrupación se le asocia manualmente una sentencia SPARQL, la cual describe un patrón de consulta específico. El módulo de implicación textual relaciona las preguntas realizadas al sistema en lenguaje natural con las preguntas agrupadas anteriormente, con el objetivo de asociar a cada pregunta su sentencia SPARQL correspondiente. Este módulo está compuesto por deducciones que infieren la relación entre dichas preguntas y las sentencias SPARQL [11]. A continuación se describe brevemente la arquitectura de QACID: El proceso comienza con el análisis de la pregunta, con el objetivo de obtener la representación formal de la misma (se analiza morfológicamente la pregunta y se detectan sus entidades, utilizando para ello un lexicón de conceptos ontológicos). La representación formal de la consulta es procesada, con el fin de determinar las implicaciones semánticas en la misma y el conjunto de patrones que contiene en la base de conocimiento. Si el proceso termina con éxito se obtiene una sentencia SPARQL, la cual permite obtener la respuesta a dicha consulta.

PowerAqua[7] es un sistema de preguntas y respuestas, el cual toma como entrada una consulta en lenguaje natural, y es capaz de devolver respuestas procedentes de cualquier lugar de la Web Semántica. La arquitectura de PowerAqua está compuesta por tres componentes: un componente lingüístico (analiza la consulta en lenguaje natural, dando como resultado un conjunto lingüístico triple, llamado Query-Triples (QTs), que identifica las asociaciones entre el conjunto de palabras en la oración), un componente llamado PowerMap (recibe los QTs producidos por el componente lingüístico, e identifica la semántica de la palabras para poder responder a la consulta dada. Así, establece las relaciones entre los términos QTs y la semántica de la palabras), y finalmente, un componente de mezcla y de categorización de las relaciones definidas por el componente anterior (genera la respuesta final) [24].

WebQA[8] es un sistema de preguntas y respuestas expresadas en español. Este sistema utiliza el etiquetador sintáctico Freeling[9] para encontrar la categoría léxica de cada palabra de la pregunta. Además, utiliza patrones que permiten caracterizar los tipos de preguntas y de respuestas, generando reformulaciones de las consultas de los usuarios, que son enviadas como consultas a un motor de búsqueda. WebQA aplica tres técnicas para reformular las consultas, las cuales son [9]: un re-formulador según patrones preestablecidos, un re-formulador en

[6] Sistema de búsqueda de respuestas basado en ontologías, implicación textual y entornos reales.
[7] http://poweraqua.open.ac.uk:8080/poweraqualinked/jsp/index.jsp
[8] Respuesta Automática a Preguntas
[9] http://nlp.lsi.upc.edu/freeling/

base a la permutación de las dos primeras palabras, y un re-formulador en base a la conjunción de palabras claves.

SWSNL[10] está construido sobre el paradigma de la Web Semántica, y usa una ontología con el fin de almacenar su base de conocimientos. Permite que los usuarios realicen sus preguntas en lenguaje natural, cuenta con un módulo de interpretación de la consulta en lenguaje natural, donde realiza un pre-procesamiento, reconocimiento de entidades, y utiliza un modelo de análisis semántico estocástico para interpretar la consulta. Ese módulo realiza una interpretación semántica de la consulta, y genera la consulta en SPARQL. Finalmente, el módulo de resultados de la búsqueda es el encargado de obtener los resultados [18].

En particular, nosotros proponemos un Marco Ontológico Dinámico Semántico, cuyas principales características son:

- ✓ Permite usar el lenguaje natural español para realizar consultas sobre la Web.
- ✓ Es basado en un conjunto de ontologías para el procesamiento de la consulta en lenguaje natural.
- ✓ Posee un componente de aprendizaje ontológico que le permite actualizar sus ontologías.
- ✓ Realiza una expansión semántica de las consultas usando el contenido de sus ontologías.
- ✓ Explota el contenido semántico sobre la Web en procesos de razonamiento automático y de aprendizaje, con el fin de optimizar los procesos de búsqueda y de aprender sobre el uso y el contenido de la Web.

Algunos de los trabajos previos tienen algunas de esas características, pero no todas a vez (ninguno de ellos tienen todas esas cualidades juntas). Ahora bien, para lograr el proceso adaptativo en el MODS se requiere de un componente de aprendizaje que permita el proceso de adquisición de conocimiento, y de actualización del contenido de sus componentes. Este aspecto es el principal aporte propuesto alrededor del MODS, lo que le permite adaptarse a su entorno, y es su principal diferencia con respecto a los trabajos previos mencionados. En el capítulo 4 se presenta el estado de arte en Aprendizaje Ontológico.

A continuación se presenta una tabla comparativa de los trabajos relacionados con respecto al MODS.

[10] SWSNL: Semantic Web Search Using Natural Language

Tabla 1.1. Tabla comparativa de los antecedentes con respecto al MODS

	Recursos lingüísticos	**Ventajas**	**Desventajas**
MESIA	• Aries: Analizador léxico morfológico • EWN: Base de datos semántica que permite obtener sinónimos	• Integra Aries y EWN • Descompone la consulta en Sintagmas nominales, verbales, preposicionales • Análisis sintáctico para extraer los términos relevantes • Términos relevantes son expandidos por EWN	• Resultado en XML, el cual no tiene semántica. • No se puede realizar inferencias con los resultados. • Los recursos lingüísticos son estáticos.
QUACID	• Lexicón de conceptos ontológicos • Patrones de consultas en SPARQL	• Análisis Morfológico • Extrae las entidades de cada termino del lexicón de conceptos • Asocia la consulta con los patrones ya establecidos	• Resultado en SPARQL, no se puede realizar inferencias con el resultado • Consultas en SPARQL son manuales en relación al historial de la consulta • Los recursos lingüísticos son estativos
PowerAgua	• Datos en RDF • WordNet	• Obtiene los datos RDF. • Valida la semántica del dato con WordNet • Luego realiza una mezcla	• Resultado en RDF, y no se puede realizar inferencias con el resultado. • Los recursos lingüísticos son estáticos. • Solo es validos para consultas en Ingles.
SWSNL	• Ontología propia	• Reconocimiento de Nombres Propios • Análisis semántico estadísticos • Interpretación semántica	• Resultado en RDF, y no se puede realizar inferencias con el resultado • Los recursos lingüísticos son estáticos.
Propuesta MODS	• Cuenta con un lexicón, onomasticon, interpretativa propias del MODS	• Realiza análisis léxico-morfológico, sintáctico, semántico y pragmático. • Cuenta con una ontología de tarea para el proceso de interpretación y una ontología lingüística para el análisis sintáctico y semántico • Generando la consulta en lenguaje OWL • La consulta puede ser transformada, en boolena para el caso de recuperación de información, SQL para el caso de base de datos Relacionales, SPARQL y SAPRQL-DL, en el caso de Ontologías. • Cuenta con un componente de aprendizaje para la adquisición de nuevo conocimiento para la actualización de sus recursos lingüísticos	• El MODS no presenta las debilidades de los trabajos previos

1.5. Organización de la Tesis

En el capítulo 1 se hace una descripción del problema a estudiar, y se plantean los objetivos y el estado de arte en cuanto a que trabajos se han desarrollado en el ámbito de la interpretación de consultas en lenguaje natural, haciendo hincapié en lo hecho para la Web. En el capítulo 2 se describe el marco teórico de la investigación, en específico, se presenta todo lo concerniente a Semántica Web, Ontologías y Semántica Ontológica. En el capítulo 3 se describe el Marco Ontológico Semántico Dinámico para la Web Semántica, en particular, en ese capítulo se describe la arquitectura propuesta. En el capítulo 4 se presenta el componente de Aprendizaje Ontológico del MODS. El capítulo 5 presenta diferentes pruebas y experimentaciones con los distintos componentes del MODS, y con el MODS como un todo, y se analizan y comparan los resultados con trabajos previos. Por último, en el capítulo 6 se presentan las conclusiones y trabajos futuros

Capítulo 2: Marco Teórico

En este capítulo introducimos la teoría que fundamenta la investigación. Se presenta las definiciones formales usadas a través del documento, la cual pretende en primer lugar familiarizar al lector con las siguientes áreas: Semántica Ontológica, Semántica Web y Aprendizaje Ontológico. Este capítulo está estructurado de la siguiente manera, en la sección 2.1., se presenta la Semántica Ontológica [40], que es una teoría que se encarga del estudio del lenguaje natural, así como sus formas de procesamiento del mismo. En la sección 2.2 se presenta la Web Semántica, que es una extensión de la actual Web, en la cual a la información se le agrega meta-información (conocimiento) que describe su contenido y significado, entre otras cosas, lo que facilita su futuro uso automatizado desde los computadores [5]. En la sección 2.3. se presenta el Aprendizaje Ontológico [26], área que integra múltiples disciplinas para facilitar la construcción y actualización de ontologías. Por último, la sección 2.4. define otros conceptos necesarios para cumplir con los objetivos planteados en este trabajo, tales como relevancia de documentos y extracción de conocimiento.

2.1. Semántica Ontológica

Antes de comenzar con la Semántica Ontológica, es necesario introducir el concepto de ontología, el cual se describe a continuación.

2.1.1. Ontologías

Existen varias definiciones del término ontología según la disciplina donde se quiera usar. En computación, en especial en el área de Inteligencia Artificial (IA), se han propuesto varias definiciones, una de ella es la propuesta por Gruber [16]: “especificación explicita de una conceptualización”.

En general, se puede decir que las ontologías proporcionan una caracterización de conceptos, y las relaciones entre ellos, en un dominio común. Ellas tienen los siguientes componentes, que sirven para caracterizar el conocimiento en un dominio [16]:

- ✓ Conceptos (llamados también clases): son las ideas básicas que se intentan formalizar, es decir, describen los conceptos de un dominio. Los conceptos pueden ser clases de objetos, métodos, planes, estrategias, procesos de razonamiento, etc. Las clases en una ontología son usualmente organizadas en taxonomías mediante el mecanismo de herencias.
- ✓ Relaciones (llamadas también propiedades de los objetos, slots[11]): representan las interacciones y enlaces entre los conceptos de un dominio. Suelen formar la taxonomía del dominio. Por ejemplo: subclase-de, parte-de, parte-exhaustiva-de, conectado-a, etc. Las ontologías usualmente tienen relaciones

[11] Termino utilizado en Protégé

binarias, donde el primer argumento es el dominio y el segundo argumento es el rango. Las relaciones binarias son algunas veces usadas para expresar los atributos de los conceptos. Los atributos se distinguen de la relación porque el rango es un tipo de dato, tal como cadena de caracteres, número, etc., mientras que el rango de la relación es otro concepto. También se puede expresar relaciones que no son binarias.

- ✓ Funciones: son un tipo concreto de relación, donde se calcula un elemento mediante una función que considera varios elementos de la ontología.
- ✓ Instancias (llamado también individuos): se utilizan para representar objetos determinados (concretos) de un concepto.
- ✓ Axiomas: son teoremas que se declaran sobre relaciones que deben cumplir los elementos de la ontología. Los axiomas son usados para verificar la consistencia de la ontología, y para inferir nuevo conocimiento.

Una forma de representar formalmente las ontologías es por medio de lógica descriptiva (DL[12] por sus siglas en inglés). Teóricamente, la lógica descriptiva se divide en dos partes: el TBox y el ABox. El TBox contiene las definiciones de conceptos y reglas, mientras que el ABox contiene las definiciones de los individuos (instancias). Básicamente, DL permite la representación en ontologías de los siguientes componentes: conceptos, relaciones e individuos.

A continuación se describe formalmente una ontología, a partir de esa definición caracterizaremos las ontologías del MODS.

Definición 1: Una ontología es una estructura $O := (C, R, A, Ax)$, que consiste de:

- ✓ Cuatro conjuntos distintos de:
  - Conceptos (C),
  - Relaciones (R)
  - Atributos (A)
  - Axiomas (Ax)
- ✓ Los conceptos se pueden estructurar/organizar en forma taxonómica (o jerarquía de conceptos) o no taxonómica.
- ✓ Las relaciones representa los vínculos entre los conceptos. Por lo general, las relaciones son del tipo binarias, donde el primer argumento de la relación se conoce como el dominio y el segundo argumento se conoce como el rango (anteriormente se explicó en qué consistía cada uno).
- ✓ Los Atributos son las características que describen a un concepto.
- ✓ Los axiomas o reglas, las cuales son definidas en un lenguaje lógico, describen las restricciones/condiciones que deben cumplir las relaciones/conceptos de una ontología dada.

Definición 2: Sea L un lenguaje lógico y Ax el conjunto de axiomas. Ese conjunto Ax es escrito usando el lenguaje L.

A continuación se ilustra con un simple ejemplo las definiciones anteriores:

[12] Description Logics

Supongamos una ontología compuesta por:

- ✓ El conjunto de conceptos $C \coloneqq \{Persona, Estudiante, Profesor, Curso\}$,
- ✓ El conjunto de relaciones $R \coloneqq \{dictaElCurso, estaInscriptoEn, subConceptoDe\}$,
- ✓ El conjunto de atributos $A \coloneqq \{tieneNombre\}$,
- ✓ Además, las relaciones entre los conceptos son:

$$\{Estudiante\ subConceptoDe\ Persona, Profesor\ subConceptoDe\ Persona,$$
$$Profesor\ dictaElCurso\ Curso, Estudiante\ estaInscriptoEn\ Curso,$$
$$Persona\ \ tieneNombre\ String\}$$

Por otro lado, el conjunto de Axiomas se define en lógica descriptiva. Dos ejemplos de axiomas serían:

- ✓ Dos conceptos siempre son disjuntos: $C \sqsubseteq \neg D$, donde C y D son dos conceptos. Por ejemplo, $Curso \sqsubseteq \neg Persona$
- ✓ Un concepto puede ser una subClaseDe de otro concepto: $C \sqsubseteq D$. Por ejemplo, $Profesor \sqsubseteq Persona$

En la Figura. 2.1. se muestra el grafo ontológico del ejemplo.

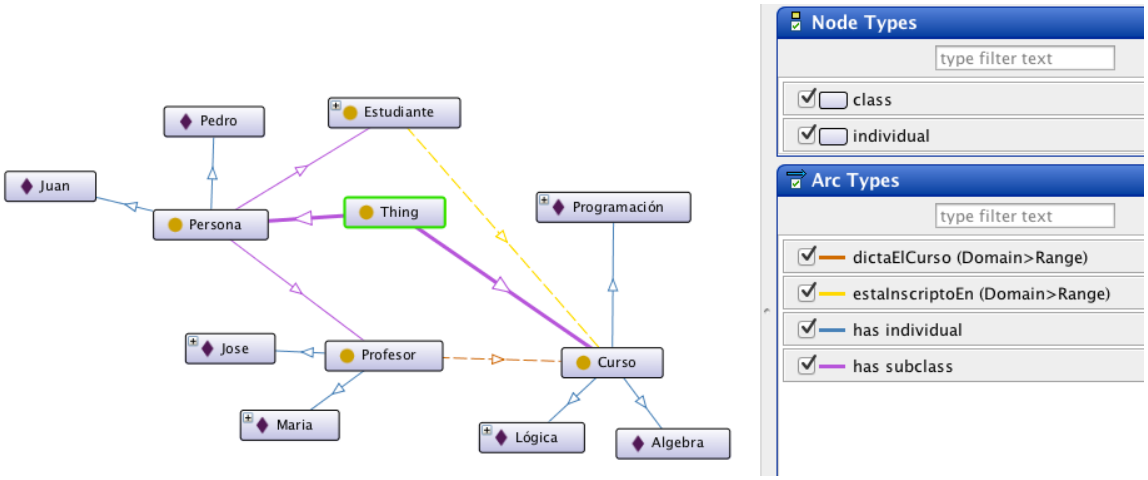


Figura. 2.1. Ontología en Protégé[13] que describe el ejemplo

2.1.2. Semántica Ontológica

La Semántica Ontológica es una teoría que se encarga del estudio del significado del lenguaje humano o lenguaje natural, así como del procesamiento del mismo [40]. Para ello, utiliza un modelo abstracto del mundo (llamada ontología) para extraer y representar los significados, los cuales son usados después para razonar y generar conocimiento.

[13] Protégé es un editor de ontología y un marco para construir sistemas inteligentes. Disponible en http://protege.stanford.edu/

Como cualquier teoría para el PLN (Procesamiento de Lenguaje Natural), la Semántica Ontológica debe considerar los procesos de generación y de análisis del significado del texto. Un método aceptado para dichos procesos es [40]:

- ✓ Describir los significados de las palabras, y separadamente,
- ✓ Especificar las reglas para combinar los significados de las palabras en oraciones y textos.

Es decir, se divide la semántica léxica (palabra) de la semántica compositiva (sentencias). En general, los dos procesos clásicos de la semántica en el PLN (el análisis y la generación de sentencias en lenguaje natural) son diferentes por naturaleza: el análisis se centra en la resolución de ambigüedades, mientras que la generación trata de resolver la sinonimia[14] de la selección léxica.

En la Semántica Ontológica, la representación del significado de un texto (TRM Text Meaning Representation) es obtenido por:

- ✓ Establecer los significados léxicos de las palabras individuales y de las frases que comprenden el texto.
- ✓ Eliminar las ambigüedades de esos significados.
- ✓ Combinar esos significados en la estructura semántica final del texto.

Un ejemplo de implementación de un TRM es propuesto en [27], en el que se realiza el proceso de análisis de un texto de entrada, dado en cualquiera de los idiomas soportados. El sentido del texto de entrada, derivado del análisis de su información léxica, sintáctica, semántica y pragmática, se representa en el TMR.

En [37] muestra un ejemplo de TMR para la siguiente entrada: “El grupo Roche, a través de su compañía en España, adquirió Doctor Andreu”. El TMR generado es mostrado en la Figura 2.2.

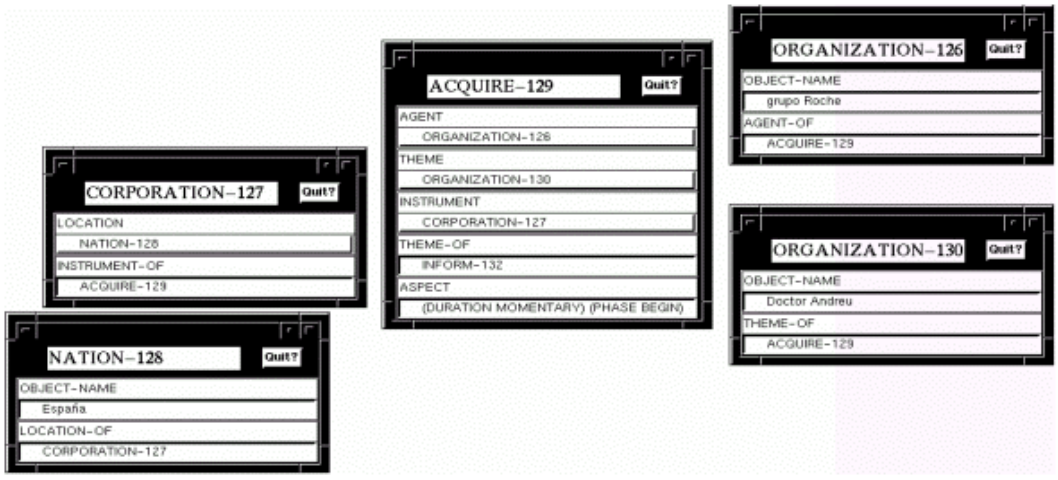


Figura. 2.2 TMR parcial de un texto de ejemplo [37]

[14] “es una relación de semejanza de significados entre determinadas palabras”. [Fuente: http://es.wikipedia.org/wiki/Sinonimia]

El significado de esta oración en TMR es descrito de la siguiente manera: el evento ACQUIRE-129 relaciona la entidad ORGANIZATION-126 denominado "grupo Roche", con la entidad ORGANIZATION-130 denominada "Doctor Andreu", a través del instrumento CORPORATION-127 ubicado en la NATION-128 llamada España. En otras palabras, el significado del texto que aparece en el TMR es: La organización llamada "grupo Roche" adquirió la organización llamada "Doctor Andreu", a través de la corporación situada en la nación llamada España.

2.2.3. Arquitectura de la Semántica Ontológica

Clásicamente, la arquitectura de la semántica ontológica es la siguiente (ver Figura 2.3, que es una adaptación de [40]):

- ✓ Un conjunto de fuentes de conocimiento, que está compuesto por: una ontología, un repositorio de hechos, un lexicón, un onomasticon [15], y elementos no semánticos (allí se encuentran los analizadores semánticos.)
- ✓ Lenguajes de representación del conocimiento, para especificar las ontologías, las unidades léxicas, etc.
- ✓ Un conjunto de módulos de procesamiento, al menos un analizador (el cual procesa datos de entrada, que pueden ser textos, preguntas etc.) y un generador (el cual genera datos de salidas, las cuales pueden ser textos, respuestas, etc.)

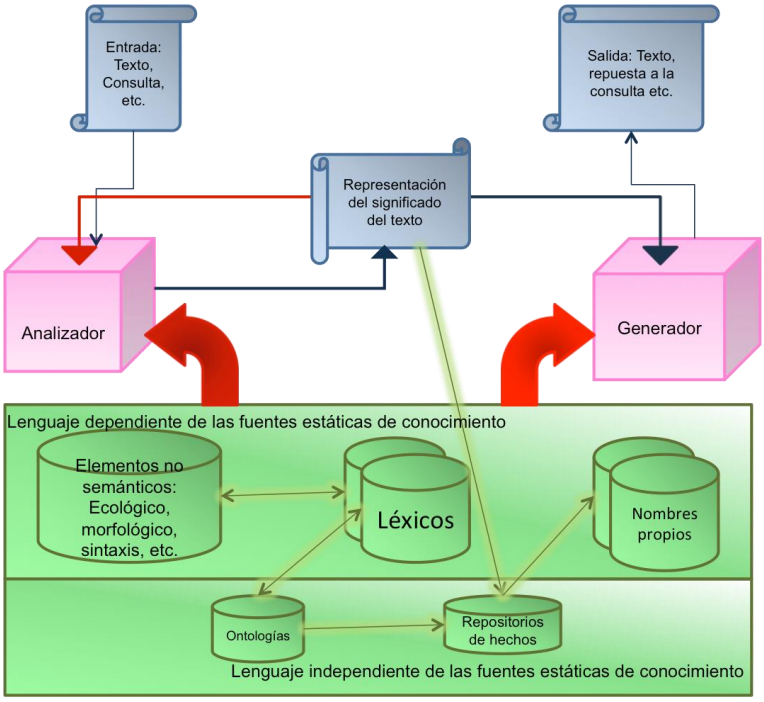


Figura. 2.3. Arquitectura de la semántica ontológica. (Adaptada de [40])

[15] Onomasticon: Relativo a los nombres, y especialmente a los propios [40]

Como hemos venido diciendo, la Semántica Ontológica se centra en el proceso de análisis y de generación del texto. En específico, el modelo de Análisis de la Semántica Ontológica propuesto en [40] es mostrado en la Figura 2.4, el cual se explica con más detalle por ser el componente que utilizaremos en este trabajo de investigación (el modelo de generación solo lo mencionaremos brevemente).

En el caso de la Figura 2.4, a partir de un texto a procesar, un analizador léxico ecológico[16] se encarga de transformar las secuencias de símbolos en unidades léxicas y, usa un conjunto de reglas para resolver posibles ambigüedades léxicas de categoría. Estos analizadores se denominan etiquetadores gramaticales. Algunos ejemplos de estos etiquetadores son relax[17] (español, catalán e inglés), TreeTagger[18] (español, francés e inglés), y Brill's tagger[19] (inglés).

Los analizadores morfológicos se utilizan para describir la información de las frases a partir de los elementos básicos de la lengua (morfemas[20], palabras y oración), señalando los grupos de elementos que funcionan como un todo, lo que permite identificar las entidades del texto a analizar. En particular, la morfología trata las palabras, tomadas independientemente de sus relaciones en la oración, y estudia sus formas. La información morfológica que proporciona una palabra incluye datos sobre su flexión (genero, número, persona,…), derivación (sufijos, prefijos,…) y composición (palabras simples, palabras compuestas). Asimismo, es objeto del estudio morfológico la categoría gramatical de las palabras (nombre, verbo, adverbio,…). La correcta identificación de unidades morfológicas es esencial para cualquier proceso posterior [46]. Si se combina el análisis morfológico con el léxico, se puede obtener información morfológica más completa sobre las unidades léxicas ya desambiguabas. Algunos analizadores morfológicos son maco+ (español, catalán e inglés) y PC-KIMMO (inglés) [46].

Los analizadores sintácticos agrupan los constituyentes de las frases, tomando como entrada el resultado del proceso de análisis morfológico. Ellos permiten extraer componentes más grandes de las palabras de un corpus textual[21], e identificar sintagmas[22] que se pueden agrupar posteriormente en oraciones. Algunos analizadores sintácticos son SUPP, tacat y Conexor [46].

A partir de la estructura generada por el analizador sintáctico, el analizador semántico se encarga de extraer el significado o sentido de la oración, y genera una estructura lógica. Una de las tareas claves de la interpretación semántica consiste en considerar qué combinaciones de significados de palabras individuales son posibles a la hora de crear un significado coherente de la oración, lo que puede reducir el número de posibles significados para cada palabra de una oración determinada.

[16] El análisis léxico morfológico, en Semántica Ontológica se llama ecológico
[17] Relaxation Labelling Based Tagger.
[18] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
[19] http://www.cs.jhu.edu/%7Ebrill/
[20] Morfema, signo lingüístico mínimo en que pueden descomponerse las palabras de una lengua. Constituye la unidad mínima del análisis morfológico o gramatical. [40]
[21] Colección de textos seleccionados que sirven de base para realizar un análisis terminológico. [40]
[22] Un sintagma es un grupo de palabras que realizan la misma función sintáctica [40].

El procesamiento pragmático interpreta la oración completa dentro de su contexto o discurso. Por ejemplo:"¿Sabe qué hora es?" se reinterpreta como petición de hora.

En la Figura 2.5 se muestra el proceso de generación, el cual se inicia con la entrada del TMR, y se encarga de generar un texto en lenguaje natural haciendo el proceso inverso del análisis.

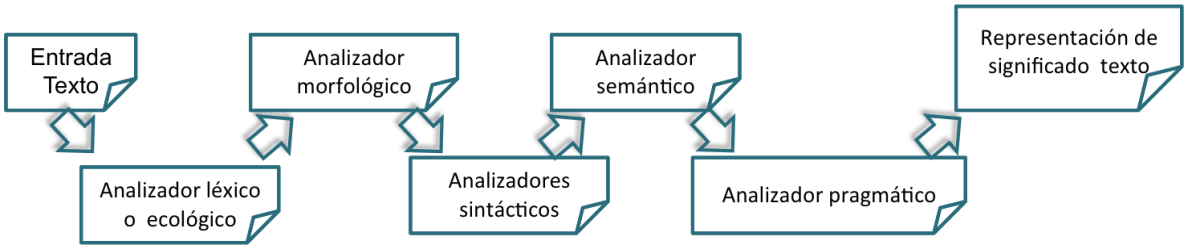


Figura. 2.4. Una visión esquemática del módulo de análisis de un sistema de PLN. (Adaptada de [40])

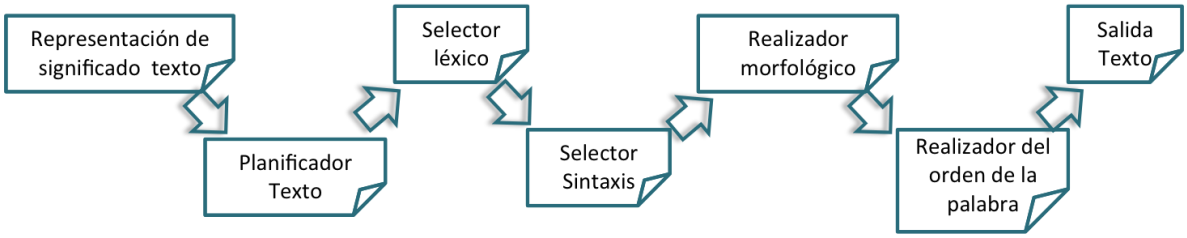


Figura. 2.5. Una visión esquemática del módulo de generación de un sistema de PLN. (Adaptada de [40])

Para el proceso de análisis se requiere de un conjunto de recursos lingüísticos, los cuales no son componentes de un sistema de procesamiento de lenguaje natural como tal, pero son fundamentales para él. El término de recursos lingüísticos se refiere al "conjunto de datos de un lenguaje, en formato legible por máquinas" [42]. Los recursos tienen como objetivo establecer explícitamente una o varias relaciones entre las palabras de un idioma específico. El uso de recursos lingüísticos mejora el desempeño del proceso de análisis, debido a que describen las palabras en forma morfológica y semánticamente correctas. A continuación se mencionan brevemente algunos de los recursos lingüísticos que serán utilizados en esta investigación:

- ✓ Lexicón: es la colección de palabras válidas de un lenguaje (en este caso, el español), que son indexadas desde el lexema[23] de la palabra, y describe todos sus posibles usos [40, 41]. Las palabras se agrupan en categorías [40]: sustantivos (pronombres y nombres, para denotar cosas), verbos (para denotar

[23] Lexema. "Unidad léxica mínima que carece de morfemas, como luz, o resulta de haber prescindido de ellos, como bland en ablandar, y que posee un significado semántico y definible por el diccionario, y no gramatical como el morfema" http://www.wordreference.com/definicion/lexema

sucesos), adjetivos (para modificar sustantivos), y adverbios (para modificar verbos).

- ✓ Onomasticon: En él se describen los nombres propios utilizados para denominar personas, organizaciones, lugares, entre otras cosas, según un dominio especifico.
- ✓ Corpus y diccionarios on-line: Los corpus son una colección de textos en lenguaje natural. Ellos, en conjunto con los diccionarios electrónicos en formato legible, permiten extraer información morfológica de las palabras.

Otro concepto vinculado al procesamiento del lenguaje Natural es:

- ✓ Lematizador: Es una parte del procesamiento lingüístico que determina el lema[24] de una palabra o token. Para ello, se utiliza un algoritmo de "stemming" que permite la normalización lingüística, tal que las diferentes formas que puede adoptar una palabra son reducidas a su forma canónica, por ejemplo: perr para las palabras perros, perras, perrito, etc. Nuestro sistema utiliza una versión del algoritmo de Porter en español, el cual es utilizado para el stemming [48].

2.2. Web Semántica

Es un hecho que la Web Semántica será el estándar de la Web en los próximos años, debido a que las personas que antes eran solo lectoras, se están convirtiendo en productoras de información. Para ello, se usan ontologías en un lenguaje estándar para representar la información de la Web Semántica.

Antes de definir qué es Web Semántica, se comenzará con la definición de semántica: "El estudio del significado de los signos lingüísticos y de sus combinaciones, desde un punto de vista sincrónico o diacrónico"[25], por lo tanto, la semántica representa el significado de las palabras, lo que permite sus futuros usos de formas más eficientes. En cuanto a la Web Semántica, en la literatura se encuentra:

- ✓ Para Berners-Lee (el creador de la WWW), la Web Semántica "es una extensión de la actual Web, en la cual la información se da con un significado bien definido, lo que facilita que los computadores y las personas trabajen en cooperación…" [5]. También define la Web Semántica como "una red de datos que pueden ser procesados directa o indirectamente por las máquinas" [5].
- ✓ En [19] definen la Web Semántica como una Web de datos descritos y enlazados de tal manera de establecer un contexto o semántica que se adhiere a construcciones gramaticales y lingüísticas bien definidas.

La definición oficial que W3C de la Web Semántica es [17]: "La Web Semántica es una Web extendida, dotada de mayor significado, en la que cualquier usuario

[24] Lema: "forma de citación de una palabra (por ejemplo, el lema de leíamos es leer" [30]
[25] Definición tomada de la Real Academia Española

en Internet podrá encontrar respuestas a sus preguntas, de forma más rápida y sencilla, gracias a una información mejor definida. Al dotar a la Web de más significado y, por lo tanto, de más semántica, se pueden obtener soluciones más eficientes a los problemas habituales de búsqueda de información... Esta Web extendida basada en significados se apoya en lenguajes universales, para convertir el acceso a la información en una tarea fácil y motivante"[26].

Para lograr que la Web extendida tenga un mayor significado se usan Ontologías, que es uno de los pilares de la Web Semántica. Por ejemplo, el término "Universidad de Los Andes" solo lo conocen las personas que de alguna manera han tenido alguna relación con la Universidad, pero para las computadoras es solo un término. Ahora bien, si se describe el término expresando que la Universidad de Los Andes es una Universidad, que está compuesta por tres núcleos, donde se imparten estudios de educación superior, se realizan investigaciones, etc., y se representa esa información en un formato que las máquinas puedan entender, es decir en una ontología, las máquinas podrán interpretar ese término.

La arquitectura básica de la Web Semántica se muestra en la Figura 2.6., la cual está compuesta de capas. En nuestro trabajo doctoral nos centraremos en las siguientes capas: la capa RDF+rdfshema, la capa de Ontologías y la capa Lógica. Por consiguiente, estas son las que explicaremos a continuación:

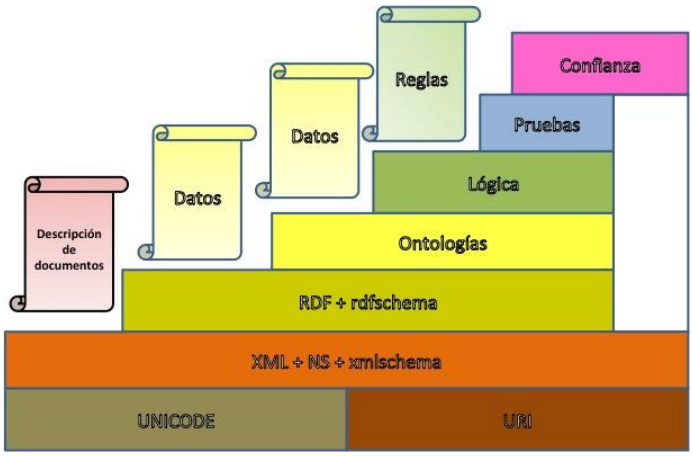


Figura. 2.6. Arquitectura de la Web Semántica. (Adaptada de [Fuente: http://www.w3.org/2000/Talks/1206-xml2k-tbl/slide10-0.html]

- ✓ RDF[27] + RDF schema[28]: esta capa define el lenguaje universal con el cual podemos expresar la información en la Web Semántica. RDF es un lenguaje

[26] http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica

[27] RDF (Resource Description Framework) fue desarrollado por W3C como un lenguaje para describir recursos de la Web, siguiendo el formalismo de representación de las redes semánticas.

[28] RDF Schema fue elaborado como una extensión de RDF con primitivas basadas en marcos. Este lenguaje permite representar conceptos, taxonomías de conceptos y relaciones binarias. RDF Schema es la combinación de ambos lenguajes, RDF y RDF Schema, se conoce como RDF(S). Se han creado motores de inferencia (integrados en interfaces de acceso a ontologías como Jena, Sesame, Oracle RDF, Boca, etc.) y lenguajes de consulta (RDQL, SeRQL, SPARQL, etc.) para manejar bases de conocimientos expresadas en dichos lenguajes. Mientras que no existe estandarización para los motores de inferencia, los lenguajes de consulta están en proceso de estandarización con el desarrollo de SPARQL

simple mediante, con el cual se definen sentencias en el formato de una 3-tupla o tripleta (sujeto+predicado+objeto). Por su parte, RDF Schema permite especificar las entidades a las que pueden aplicarse los atributos del vocabulario. Por ejemplo, en el dominio de una biblioteca se podría usar RDF para definir un vocabulario concreto (autorDe, tieneCarnetDeSocio, estaSancionado, etc.), y con RDF Schema especificar condiciones tales como que estaSancionado solo puede aplicarse a socios de la biblioteca. Así, esta capa no sólo ofrece descripción de los datos, sino también cierta información semántica.

- ✓ Ontologías: La utilización de URIs[29] para describir las propiedades de los recursos en RDF garantiza la definición única de los conceptos que representan los datos de documentos en la Web. Sin embargo, esta capa permite extender la funcionalidad de la Web Semántica agregando clases y propiedades para describir los recursos. El lenguaje OWL (Ontology Web Language) es el lenguaje estándar de la Web Semántica usado en esta capa, y el recomendado por el WC3. Este lenguaje amplía RDF con primitivas de la lógica descriptiva, permitiendo representar expresiones complejas para describir conceptos y relaciones. OWL cuenta con varios razonadores (Pellet, FaCT++, Racer) [14], que se pueden utilizar para examinar restricciones de conceptos, propiedades e instancias, y para clasificar automáticamente los conceptos de forma jerárquica.
- ✓ Lógica: está compuesta por un conjunto de axiomas y reglas de inferencia que los agentes (computacionales o humanos) podrán utilizar para relacionar y procesar la información. Estas reglas ofrecen el poder de deducir nuevas sentencias a partir de los datos y estructuras que están descritos en las capas XML y RDF, usando además las relaciones entre esos datos y estructuras definidas en la capa ontológica. La expresividad de RDF y RDFS para modelar ontologías completas es muy limitada: RDF/RDFS carecen de soporte para tipos de datos primitivos, carecen de poder expresivo para representar axiomas (no hay negación, implicación, cardinalidad, cuantificación), no es posible definir propiedades de propiedades con ellos (transitividad, simetría, etc.), no permiten especificar condiciones necesarias y suficientes para establecer la pertenencia a una clase, entre otras cosas. Esta capa se apoya en OWL, la cual tiene mayor expresividad que RDF y RDFS, con una semántica formal basada en lógica descriptiva, e importantes primitivas para la descripción de clases y propiedades, como son: relaciones entre clases (por ejemplo. complemento, disjunta), cardinalidad de propiedades (por ejemplo. mínimo dos, exactamente uno), igualdad de clases, propiedades de las relaciones (por ejemplo. simetría, transitividad), entre otras.

La Figura 2.7. muestra una arquitectura por capas alternativa propuesta últimamente, las diferencias que nos interesan resaltar para nuestro proyecto, con respecto a la arquitectura original (Figura 2.6.), son las siguientes [3]:

[29] Uniform Resource Indentifier

1. La capa ontológica se centra en la integración de OWL con un lenguaje basado en reglas para lograr mayor expresividad. Esta capa usa los siguientes lenguajes:
   - ✓ Para la descripción de los conceptos y relaciones se usa OWL.
   - ✓ Para la descripción de las Reglas se usa el lenguaje basado en reglas (RIF/SWRL[30]).
   - Para la consulta se usa el lenguaje de consulta SPARQL.
2. DLP[31] es la intersección de OWL y la Lógica de Horn[32], lo que permite incorporar todas las cláusulas de Horn de primer orden en el análisis en lógica descriptiva implícito en OWL, enriqueciendo así el procesamiento lógico computacional.

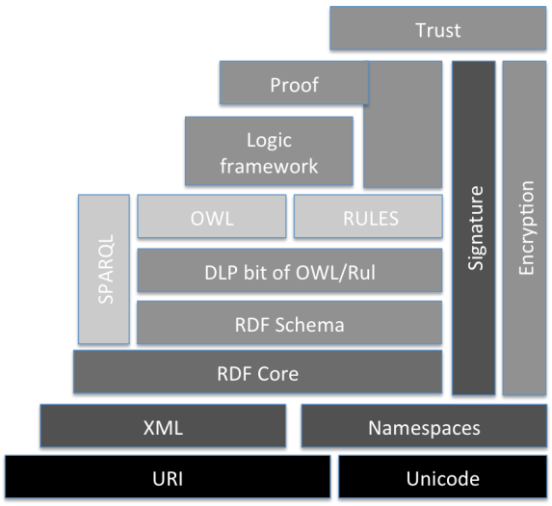


Figura. 2. 7. Arquitectura alternativa de la Web Semántica. [3]

2.3. Aprendizaje Ontológico

El aprendizaje ontológico, o aprendizaje de ontologías, facilita la construcción de ontologías para la Web Semántica, por la posibilidad de actualizarlas posteriormente. Construir y ensamblar manualmente una ontología ha sido un significativo cuello de botella para el Ingeniero de Conocimiento[33]. La dificultad radica a nivel del conocimiento que se requiere para su diseño, el tiempo y esfuerzo necesario, los potenciales errores en el proceso de diseño (por ejemplo, inconsistencias, etc.), entre otras cosas, lo que ha dado origen al área de Aprendizaje de Ontologías (AO). El proceso de aprendizaje de ontologías tiene

[30] A Semantic Web Rule Language, el cual, extiende el conjunto de axiomas de OWL para incluir reglas condicionales (Cláusulas de Horn)
[31] Description logic programs
[32] Cláusula de Horn: "Horn llegó a la conclusión que para realizar un inferencia correcta, y así eliminar las ambigüedades, las cláusulas solo debían de tener una conclusión. Es decir, debían de ser de la forma: B<-A_1,A_2,…,A_n, (una única conclusión B). Por lo tanto, si A_1,A_2,…,A_n son ciertas se pueden inferir que B es cierta". [31]
[33] Un ingeniero de conocimiento u ontológico, es alguien que investiga un dominio concreto, aprende que conceptos son los más importantes en ese dominio, y crea una representación formal de los objetos y relaciones del dominio.

que ver con la extracción de elementos ontológicos[34] a partir de diferentes fuentes, para construir o actualizar una ontología existente de manera significativa[35].

El aprendizaje ontológico se define como el conjunto de métodos utilizados para construir, enriquecer o adaptar una ontología existente de forma semiautomática (es decir, con intervención humana) o automática, utilizando fuentes de información heterogéneas. En este proceso se utiliza información estructurada y semi estructurada de diferentes fuentes para extraer conocimiento.

El aprendizaje ontológico implica un conjunto de disciplinas complementarias, en particular la ingeniería ontológica (*Ontological Engineering*) [14] y el aprendizaje de máquina (*Machine Learning* o aprendizaje automático) [22]. La ingeniería ontológica se refiere al conjunto de actividades que tienen que ver con el proceso de desarrollo de ontologías, los métodos y metodologías para construir ontologías, y las herramientas y lenguajes que las soportan. El aprendizaje de máquina consiste en la aplicación de algoritmos capaces de generalizar conocimiento a partir de una información suministrada. El modo semi-automático del aprendizaje ontológico se lleva a cabo bajo el paradigma "*modelamiento cooperativo equilibrado*" [38], que define un esquema de interacción coordinada entre el modelador humano y los algoritmos de aprendizaje para la construcción de ontologías.

Actualmente existen varios métodos de aprendizaje de ontologías. Alexander Maedche y Steffen Staub distinguen los siguientes enfoques de aprendizaje de ontologías [26]:

1. Aprendizaje de ontologías a partir de textos[36] (basado en corpus[37]): En este enfoque se aprenden conceptos y relaciones utilizando textos. El objetivo es extraer conocimiento del texto. Por lo general, la información que se extrae de los textos es la siguiente: las entidades (que son las unidades básicas del texto), las relaciones que se establecen entre las entidades, y los eventos en los que están implicadas las entidades. Los tipos de técnicas de aprendizaje para este caso se clasifican en [26]:
   1.1. Extracción basada en patrones: las relaciones y conceptos son reconocidos como secuencias de palabras en el texto que siguen un determinado patrón.
   1.2. Reglas de asociación: usadas para descubrir relaciones no-taxonómicas entre conceptos, para lo cual utilizan jerarquías de conceptos como conocimiento y estadísticas de ocurrencia de términos en el texto.
   1.3. Agrupamiento conceptual. Un conjunto de conceptos son tomados como entrada y luego son agrupados de acuerdo a la distancia semántica entre ellos (todos los conceptos cuya distancia semántica es menor que el umbral predefinido pertenecerán al mismo grupo). Una forma de calcular la

[34] Son considerados elementos ontológicos, conceptos/palabras, relaciones, funciones, axiomas, instancias.
[35] Significativa quiere decir que se estable una conexión con el conocimiento existente, y no islotes de nueva información.
[36] Primero debemos hacer una distinción entre el nivel lingüístico y conceptual, para luego entender el método propuesto. La distinción principal es que el conocimiento a nivel lingüístico se describe a través de términos lingüísticos, mientras que en el nivel conceptual el conocimiento se describe a través de conceptos y relaciones entre conceptos.
[37] "un corpus es una recopilación de fragmentos de una lengua que se seleccionan y se ordenan según un criterio lingüístico con la finalidad de ser utilizado como una muestra de la lengua o de una variedad de lengua." [39]

distancia entre conceptos es basado en las funciones sintácticas que los términos asociados a tales conceptos desempeñan en el texto.

1.4. Poda ontológica. El objetivo de la poda es obtener una ontología de dominio usando una ontología base y un corpus. Ella permite ajustar una ontología. Para ajustar la ontología se parte de bases de datos terminológicas del dominio, de las cuales se extraen los conceptos relevantes del dominio, los cuales se incorporan en una ontología genérica. Para finalizar el proceso de poda, se usan textos generales como específicos al dominio, para eliminar aquellos conceptos de la ontología genérica, basados en que tan frecuentes son esos conceptos en los textos específicos. [25].

2. Aprendizaje de ontología desde instancias. Instancias particulares, por ejemplo aquellas que son tomadas de un texto, son usadas para generar ontologías. Esto, debido a que los textos generalmente describen un dominio, por lo tanto pueden ser usadas técnicas de extracción de conceptos de los textos, para construir/actualizar ontologías.
3. Aprendizaje de Ontologías desde esquemas: esquemas de bases de datos, modelos entidad-relación, esquemas XML, etc., son usados para generar ontologías por procesos de re-ingeniería. Por ejemplo, la generación de ontologías a través de mapeos, es decir, usando XML u otro lenguaje declarativo, el cual describe el emparejamiento entre un esquema de base de datos y una ontología.

A continuación se describen, en forma general, algunos de los sistemas de aprendizaje de ontologías. Estos sistemas se caracterizan por usar diferentes técnicas de aprendizaje y diversas fuentes de información.

- ✓ ONTOLEARN[38]: Es un sistema que permite enriquecer una ontología de dominio con conceptos y relaciones. Usa aprendizaje de máquina (AM)[39], el cual ofrece algunas técnicas para el descubrimiento de conocimiento (patrones) basadas en cadenas de markov[40][35].
- ✓ LEXTER: Utiliza una máquina de estados finitos para obtener los sintagmas nominales[41] máximos. Una vez extraídos estos sintagmas, los divide en sub-sintagmas, para obtener los que aparecen en un corpus en situaciones no ambiguas [8].
- ✓ TEXT-TO-ONTO: Su principal objetivo consiste en el aprendizaje de relaciones taxonómicas y no-taxonómicas. Se basa en un análisis estadístico y un analizador sintáctico de textos en francés, más reglas de asociación y de poda/acotación. Aprende desde textos libres, semi-estructurados, desde otras ontologías, y desde bases de datos. El resultado del proceso de aprendizaje es una ontología de dominio que contiene conceptos específicos del dominio bajo

[38] http://www.gabormelli.com/RKB/OntoLearn_System
[39] El principal objetivo del Aprendizaje de Máquina (ML por sus siglas en Inglés) es el desarrollo de sistemas que puedan cambiar su comportamiento de manera autónoma basados en su experiencia.
[40] Una cadena de Márkov es una serie de eventos, cuya probabilidad de que ocurra un evento depende del evento inmediato anterior. En efecto, las cadenas de este tipo tienen memoria. "Recuerdan" el último evento y esto condiciona las posibilidades de los eventos futuros.
[41] Un sintagma nominal (SN) designa alguno de los participantes en la predicación verbal. Está constituido por un nombre (sustantivo) o adjetivo. También se le llama frase nominal (FN).

estudio. Todo el proceso es supervisado por el ingeniero ontológico [25].

2.3.1 Aprendizaje de relaciones no-taxonómicas

El aprendizaje de relaciones no-taxonómicas consiste en el descubrimiento de verbos relacionados con el dominio, la extracción de conceptos que están relacionados no taxonómicamente, relaciones de marcado (por ejemplo, las etiquetas puestas a un producto, que hablan sobre el producto), entre otras. Así, las relaciones no-taxonómicas se refieren a cualquier relación entre conceptos (excepto la relación "es_un"), tales como sinónimos, merónimos[42], antónimos, etc.

Un grupo de relaciones de interés para este trabajo son:

- ✓ *Relación de equivalencia*: establece la igualdad o equivalencia entre dos conceptos aparentemente diferentes. Esta relación suele aparecer en el lenguaje natural con expresiones parecidas a las siguientes: "A equivale a B", "A es igual a B".
- ✓ *Relación de dependencia:* es un tipo especial de relación de asociación a través de responsabilidades, parentesco, propiedades, etc. Esta relación suele aparecer en el lenguaje natural en las formas siguientes: "A está asociado a B", "A es responsable de B", "A depende de B".
- ✓ *Relación topológica:* describe la distribución espacial de conceptos físicos, y las interconexiones entre esos conceptos. Esta relación suele aparecer en el lenguaje natural en las formas siguientes: "A está a la derecha de B", "A está encima de B", "A está debajo de B", "A está dentro de B", "A contiene a B", "A está conectado a B", "A está al lado de B".
- ✓ *Relación causal:* Este tipo de relación describe como, dados unos estados o acciones, se induce a otros estados o acciones. Esta relación suele aparecer en el lenguaje natural de la siguiente manera: "A es la causa de B", "A necesita a B", "A activa a B".
- ✓ *Relación funcional:* Esta relación describe las acciones y las posibles consecuencias de las mismas. Esta relación suele aparecer en el lenguaje natural en la forma siguiente: "A permite a B".
- ✓ *Relación cronológica:* Esta relación se conoce también como relación temporal, y describe la secuencia de tiempo en la que ocurren eventos. Esta relación suele aparecer en el lenguaje natural en las formas siguientes: "A ocurre antes de B", "A ocurre después de B", "A y B ocurren simultáneamente", "A ocurre durante B", "A comienza cuando B termina".
- ✓ *Relación de similaridad:* establece que conceptos son iguales o análogos, y en qué medida. Esta relación suele aparecer en el lenguaje natural en la forma siguiente: "A es similar a B"
- ✓ *Relación condicional:* define las condiciones en las cuales ciertas cosas tienen lugar. Esta relación suele aparecer en el lenguaje natural en la forma siguiente: "A tiene como condición B", "A está condicionada por B", "Si A entonces B".
- ✓ *Relación de propósito:* Esta relación establece el por qué y el para qué de los conceptos. Esta relación suele aparecer en el lenguaje natural en las formas

[42] Se denomina merónimo a la palabra cuyo significado constituye una parte del significado total de otra palabra.

siguientes: ”A ocurre con la finalidad de B”, “A ocurre para B”, ”A nace con el propósito de B”.

2.3.2 Aprendizaje de relaciones taxonómicas

Las relaciones taxonómicas permiten estructurar los conceptos dentro de categorías, con el fin de facilitar su búsqueda, reusó y entendimiento. El aprendizaje de relaciones taxonómicas consiste en el descubrimiento de relaciones taxonómicas, en particular las del tipo “es_un”. El grupo de relaciones de interés en este trabajo a aprender son [14]:

- ✓ *Subclase:* Un concepto C1 es subclase de otro concepto C2 si y solo si toda instancia de C1 es también instancia de C2. $C_1 \sqsubseteq C$
- ✓ *Partición:* Una partición de un concepto C es un conjunto de subclases de C que son disjuntas entre sí pero que cubren a C, es decir, no hay casos de C que no sean instancias de uno de los conceptos de la partición. $C_1 \cup \ldots \cup C_n \equiv C$ donde $C_i \sqsubseteq \neg C_j, 1 \leq i \leq j \leq n$
- ✓ *Descomposición disjunta*: Una descomposición disjunta de un concepto C es cuando un subconjunto de subclases de C no cubre a todo C, es decir, que pueden haber instancias de un concepto C que no son instancias de las subclases de C. $C_1 \cup \ldots \cup C_n \sqsubseteq C$ donde $C_i \sqsubseteq \neg C_j$
- ✓ *Descomposición exhaustivas*: Una descomposición exhaustiva de un concepto C es un conjunto de subclases de C que cubren a todo C, es decir, no puede haber instancias del concepto C que no sea instancia de algunas de las descomposiciones $C_1 \cup \ldots \cup C_n \equiv C$

2.4. Otros conceptos importantes vinculados al trabajo

Las tres primeras secciones son los conceptos claves para la definición del Marco Ontológico semántico Dinámico (MODS), el cual tiene como fin interpretar las consultas a la Web en lenguaje natural. En el MODS existe un componente que actualiza las ontologías con nuevo conocimiento, generado desde la información que se recupera en la Web, extraído desde la información contenida en las páginas Web recuperadas, diccionarios online, etc. Ese componente requiere la definición de los siguientes conceptos.

2.4.1 Relevancia de Documentos

Se define la relevancia de un documento para una consulta, en cuanto el mismo responde a la consulta. Es decir, se trata de determinar la similitud entre la consulta y el documento. Para determinar esa similitud se usan distintos modelos en la literatura, siendo las más importantes el booleano, el vectorial, el probabilístico, los basados en lógica difusa, redes neuronales, o redes bayesianos

[28]. Nos centraremos en el Modelo Vectorial, por ser el que será utilizado en este proyecto.

El modelo vectorial es ampliamente usado en operaciones de Recuperación de Información, así como también en operaciones de categorización automática, filtrado de información, etc. Este modelo representa la consulta y los documentos mediante vectores [7]. Así, un vocabulario de tamaño t definirá un espacio $t - dimensional$, tal que un documentos d_j es representado por un vector: $d_j = (w_{1j}, w_{2j} \dots w_{tj})$, donde w_{ij} es una variable binaria que indica si la palabra/termino i es usado por el documento j. Paralelamente, una consulta q es representado como un vector $q = (w_{1j}, w_{2j} \dots w_{tq})$.

El modelo vectorial propone determinar el grado de similitud entre los documentos de una colección dada y las consultas, mediante algún criterio que muestre la mayor o menor cercanía entre sus vectores. Una forma de cuantificar el nivel de cercanía entre los vectores es mediante el coseno del ángulo que forman los vectores (ese valor será mayor entre más cercanos estén entre si ambos vectores, y menor entre más alejados estén entre sí).

En [32] se definen dos principios esenciales de este modelo, que son:

- ✓ *Equiparación parcial*, capacidad del sistema para ordenar los resultados, basado en el grado de similitud entre cada documento recuperado y la consulta.
- ✓ *Ponderación de los términos de los documentos y de la consulta*, el cual consiste en dar un valor real a los términos que reflejen su importancia en el documento y en la consulta.

Para cumplir con los principios mencionados, se deben realizar los siguientes procesos:

- ✓ *Análisis de frecuencia*, el cual consiste en la contabilización del número de ocurrencias de los términos que se encuentra en la consulta y en los documentos recuperado,
- ✓ *Luego se obtienen los pesos TF-IDF*[43], que consiste en el cálculo de la importancia de un término para discriminar un documento. En este trabajo se usa la técnica de asignación de pesos, donde a cada término se le asigna un peso calculado en función del valor inverso de su frecuencia de aparición en el conjunto de documentos de la colección [28, 32]. Para hallar esos pesos se debe previamente definir la frecuencia inversa (IDF), la cual es calculada como [32]:

$$IDF_i = log_2(\frac{N}{DF_i}) \qquad (2.1)$$

tal que *TF-IDF* es

$$TF - IDF_{ij} = TF_{ij} * IDF_i \qquad (2.2)$$

Donde:

[43] En inglés: Term Frequency-Inverse Document Frequency

- N = número total de documentos de la colección
- DF_i= número de documentos en los que aparece el término i
- TF_{ij} = Frecuencia de aparición del término i en el documento j

✓ *La frecuencia, como los pesos, son usados para el cálculo de la similitud*. Para medir la similitud entre un documento dj y una consulta q, se va emplear la siguiente formula [7]:

$$sim(d_j, q) == \frac{\sum_{i=1}^{t}(TF - IDF_{ij} \times TF - IDF_{iq})}{\sqrt{\sum_{i=1}^{t} TF - IDF_{ij}^2} * \sqrt{\sum_{i=1}^{t} TF - IDF_{iq}^2}} \qquad (2.3)$$

Es de hacer notar que el modelo vectorial tiene como limitante la sensibilidad semántica, es decir que documentos con contexto similares pero con diferente vocabulario no serán asociados, dando falsos negativos. Para resolver esta limitante la consulta debe pasar por un proceso de interpretación y expansión, tal que se le añade a la consulta aquellos términos relacionados que pueden utilizarse para expresar la misma idea o concepto.

Otra medida que nos interesa en este trabajo es la *Precisión*, que es una medida que indica la calidad de lo recuperado, la cual se define como la proporción de los documentos recuperados que son relevantes. Para su cálculo se utiliza la siguiente fórmula [28]:

$$Precisión = \frac{(cantidad\ de\ documentos\ relevantes\ recuperados)}{(cantidad\ de\ documentos\ recuperados)} \qquad (2.4)$$

2.4.2 Extracción de conocimiento

En el ámbito de la extracción de conocimiento, recientemente ha surgido un sub campo basado en ontologías (Ontology-Based Information Extraction (OBIE)), donde las ontologías son usadas para el proceso de extracción de la información, y la salida es presentada generalmente también a través de ontologías. En este trabajo nos enfocaremos en la extracción de información desde textos no estructurados. En [49] presentan la arquitectura básica de los Sistemas de Extracción de Información Basados en Ontologías para el procesamiento de texto no estructurado (ver la Figura 2.8).

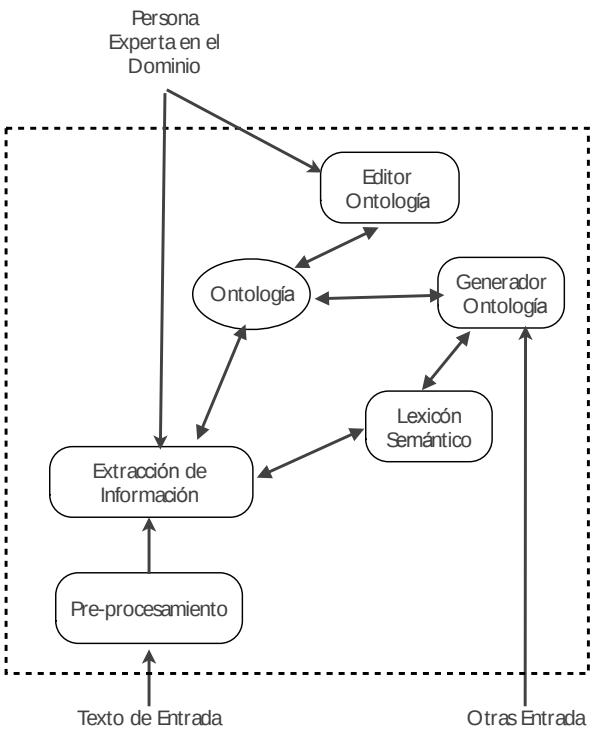


Figura. 2.8. Arquitectura General de los Sistemas de Extracción de Información Basado en Ontologías [41]**.**

La Figura 2.8. muestra los diferentes componentes de estos sistemas, aunque muchos de los sistemas no contienen todos los componentes de esta arquitectura. La descripción de esos componentes es la siguiente:

- ✓ *Pre-procesamiento:* la entrada al sistema es un texto que pasa por el componente de pre-procesamiento, el cual convierte el texto a un formato que pueda ser manejado por el módulo de extracción de información. Parte de las tareas que hace es eliminar todas las etiquetas html, caracteres especiales, etc., y convertir la entrada en un archivo texto.
- ✓ *Extracción de Información* (EI): Este módulo tiene como objetivo identificar y extraer información automáticamente desde textos en lenguaje natural. Algunas de las tareas que realiza son: reconocimiento de las Entidades (nombres de personas, lugares, organizaciones, etc., conocido en inglés como Named Entity Recognition.), clasificación de subcadenas de textos que hacen referencias a entidades del mundo real, reconocimiento de las funciones semánticas de los componentes de una oración, reconocimiento de las relaciones entre entidades, clasificación de las relaciones entre dos o más entidades según su función semántica. Puede ser implementado usando cualquiera de las siguientes métodos: reglas lingüísticas representadas por expresiones regulares, listas/diccionarios especializados de nombres propios (de personas, compañías, lugares, etc.), técnicas de clasificación, arboles sintácticos parciales, analizadores de tags HTML/XML, entre otros [41, 39 y 43].

- ✓ *Lexicón semántico para el lenguaje*, cuyo objetivo es definir las palabras válidas del lenguaje, descritas semánticamente (sus posibles formas y usos, etc.). Un ejemplo es WordNet.
- ✓ *Sistema de Gestión de Ontologías*, el cual permite editarla, modificarla, generarla y usarla. Es usada por el sistema para generar internamente elementos de la ontología, pero también por expertos de un dominio para asistir de manera semiautomática, utilizando un editor de ontologías, al sistema en el proceso de su actualización.
- ✓ *La salida del sistema* consiste en el conocimiento extraído desde los textos, que puede estar representado usando lenguajes ontológicos, como por ejemplo OWL.

2.5. Conclusión

Los aspectos teóricos estudiados para la resolución del problema, son los siguientes: el proceso de interpretación de la consulta natural es basado en el modelo de análisis de la semántica ontológica mostrado en la figura 2.4, adaptándolo a las necesidades del MODS, donde sus recursos lingüísticos son ontologías y cuenta con una ontología lingüística para el procesamiento sintáctico y semántico; en cuanto a la formalización se basó en la WEB Semántica para la anotación, utilizándose XML, RDF, RDFS, OWL para la representación del conocimiento, y RDFS, OWL, SWRL para las inferencias; para el proceso de adaptabilidad de los componentes del MODS se utilizó de los procesos de aprendizajes de ontologías el aprendizaje de términos y el aprendizaje de relaciones taxonómicas y no taxonómica. Finalmente, se utilizó el modelo vectorial para la obtención de documentos relevantes, y la extracción de conocimientos para la extracción de entidades y relaciones.

Capítulo 3: Marco Ontológico Dinámico Semántico (MODS).

El objetivo general de este trabajo es la propuesta de un Marco Ontológico Dinámico Semántico (MODS) para la Web Semántica, con el fin de que sea un interpretador de consultas en lenguaje natural para la Web. De esta manera, se aspira lograr interpretar y formalizar las consultas en lenguaje natural enviadas a la Web. En este capítulo se describe el Marco propuesto, además se describe cada una de las ontologías utilizadas, siguiendo con la caracterización del proceso de consulta en lenguaje natural usando las ontologías. Este capítulo está organizado de la siguiente manera, en la sección 3.1 se presenta la Arquitectura del MODS, en la sección 3.2 se describen los componentes del MODS, y en la sección 3.3 se describe el comportamiento del MODS.

3.1. Marco Ontológico Dinámico Semántico.

El MODS es una propuesta novedosa, que permite el análisis y la realización de consultas en lenguaje natural en la Web Semántica. La consulta para el MODS, más que una petición de información, es un elemento cargado de información útil para formarse una idea del tipo de usuario, y aproximarse, de manera sucesiva, a una respuesta que cada vez más satisfaga las necesidades del mismo. MODS tiene el desafío de interpretar y formalizar la consulta realizada por el usuario en lenguaje natural, y refinar sus esquemas internos frente a la dinámica de la Web, para tratar futuras consultas (para eso requiere de mecanismos de adaptación). A continuación se presenta la arquitectura del MODS (ver Figura. 3.1).

De manera general, MODS transforma la consulta a un lenguaje ontológico, utilizando sus diferentes componentes: el lexicón, la ontología lingüística, la ontología de tareas y la ontología de dominio. De esta manera, MODS utiliza mecanismos de la semántica ontológica y herramientas del procesamiento del lenguaje natural para el procesamiento de las consultas de los usuarios. Una descripción más detallada de la arquitectura es la siguiente (ver Figura 3.1): la ontología de tareas modela el procesamiento de la consulta en lenguaje natural (análisis léxico-morfológico, análisis sintáctico-semántico, análisis pragmático); la ontología lingüística especifica la gramática del lenguaje español, y cuenta con una extensión de derivaciones coloquiales y con un lexicón para caracterizar al lenguaje español, que a su vez contiene un onomasticon; finalmente, la ontología interpretativa modela el conocimiento sobre el contexto específico del usuario (es una ontología de alto nivel, con especializaciones/extensiones basadas en ontologías de dominio externas al MODS). La ontología interpretativa, entre otras cosas, tiene una ontología del usuario, que describe el uso del sistema que va haciendo cada usuario, lo que permite incorporar a la consulta formal las características propias del usuario (contextualización), para intentar delimitar la respuesta de la Web. Finalmente, otro componente clave para la adaptabilidad del

MODS a la dinámica de la Web y del usuario, es el componente de aprendizaje de ontologías, cuyo fin es permitir que las ontologías evolucionen a la par con la usabilidad del sistema.

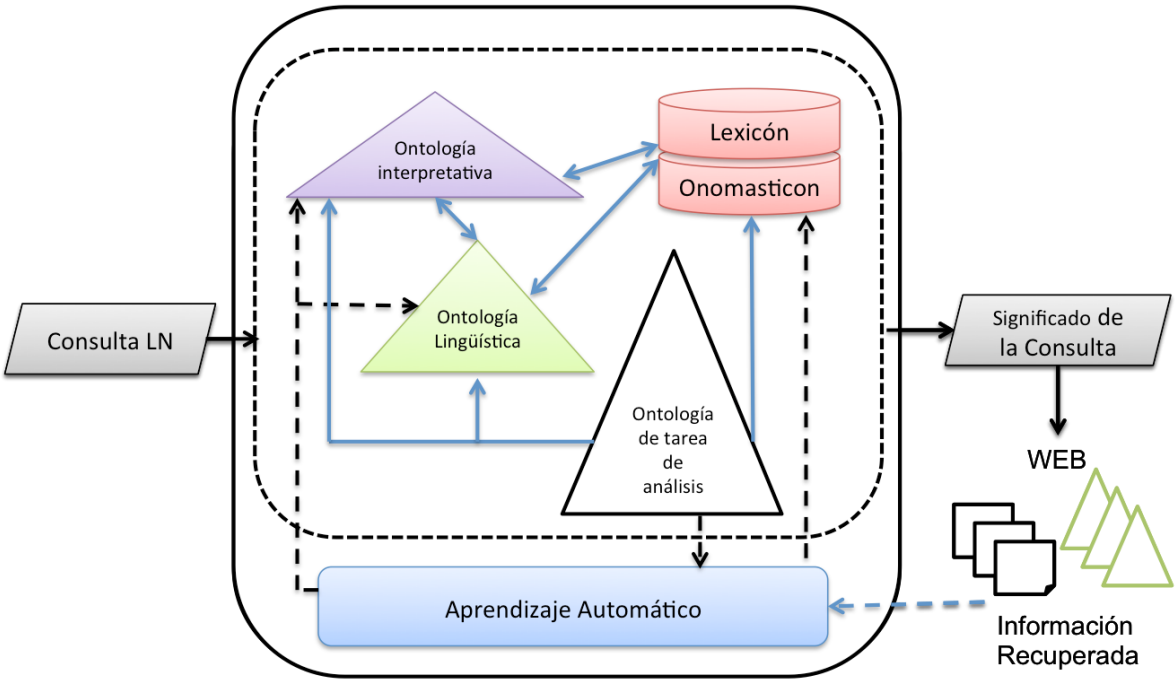


Figura. 3.1. Marco Ontológico Dinámico Semántico**.**

3.2. Descripción de los componentes del MODS

Como se puede observar en la Figura 3.1, el MODS consta de cuatro sub-marcos ontológicos, cada uno ha sido conceptualizado usando la herramienta Protégé [44], la *ontología del lexicón* que describe los términos, la *ontología lingüística* que describe la gramática del español [13], la *ontología interpretativa* que describe al dominio donde se esté usando MODS (por ejemplo Universitario), y la *ontología de tareas* conceptualiza el proceso de análisis de la consulta. Pasemos a describir a cada una:

3.2.1. Ontología del Lexicón

El lexicón del MODS caracteriza las palabras, en esta ontología de lexicón se describe la estructura morfológica de ellas. Las palabras se agrupan en categorías

[44] http://protege.stanford.edu/

(verbo, sustantivo, adverbio y adjetivo); además, se identifican por sus lemas (el lema para los sustantivos es usualmente el masculino y singular, mientras que para los verbos es el infinitivo, ver Figura. 3.2) y lexemas (raíz de la palabra).

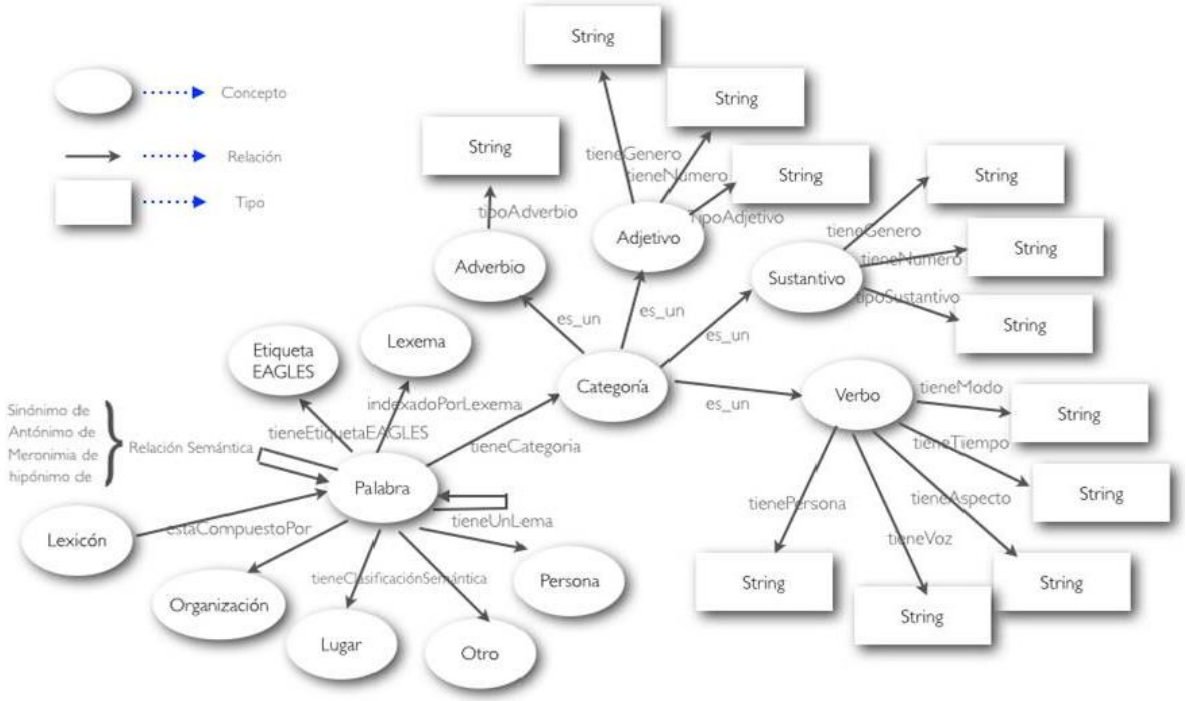


Figura. 3.2. Esquema Conceptual del Lexicón.

Por otro lado, el lexicón del MODS nos da las relaciones semánticas: sinónimos [45] (synonymy), hipónimo [46] (hyponym), antónimos [47] (antonymy), Meronimias [48] (meronymy) y la polisemia [49]. Además, usa el estándar de codificación léxica EAGLES [50] y su clasificación semántica definida por: organización, lugar, persona y otro (ver el esquema conceptual en la Figura 3.2)

A continuación, damos un ejemplo de la caracterización ontológica del conocimiento almacenado en el lexicón, para el caso de la *relación hiponimia:*

Si $B \sqsubseteq B\ (Refexividad)$, $B \sqsubset A \Rightarrow \neg A \sqsubset B\ (Antisemetría)$, $B \sqsubset A \bigwedge Z \sqsubset B \Rightarrow Z \sqsubset A(Transitividad)$, entonces B es un hipónimo de A

En el anexo A hay otros ejemplos del conocimiento guardado en la ontología de lexicón.

[45] Sinónimo: Vocablos o expresiones que tienen una misma o muy parecida significación: "esposos" y "cónyuges" son términos sinónimos. http://www.wordreference.com/definicion/sinonimos
[46] Hipónimo: palabra cuyo significado es más específico que otro, y es englobado por ese otro significado: “minuto” es un hipónimo de “tiempo”. http://www.wordreference.com/definicion/hip%C3%B3nimo
[47] antónimo: Palabra que expresa una idea opuesta o contraria a la expresada por otra palabra: "vicio" y "virtud" son palabras antónimas; el antónimo de "claro" es "oscuro".http://www.wordreference.com/definicion/ant%C3%B3nimo
[48] Meronimia: es la correspondencia léxica de la relación parte-todo. Decimos que el dedo es un merónimo de la mano [36].
[49] Polisemia: es polisémica si tiene más de un sentido relacionado entre sí [36].
[50] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Además, el MODS cuenta con un onomasticon (ver Figura. 3.3), el cual representa los nombres propios, el cual está compuesto por palabras simples y compuestas de tipo sustantivo. También usa la clasificación semántica del estándar de codificación léxica EAGLES para caracterizar a cada nombre propio: organización, lugar, persona y otro.

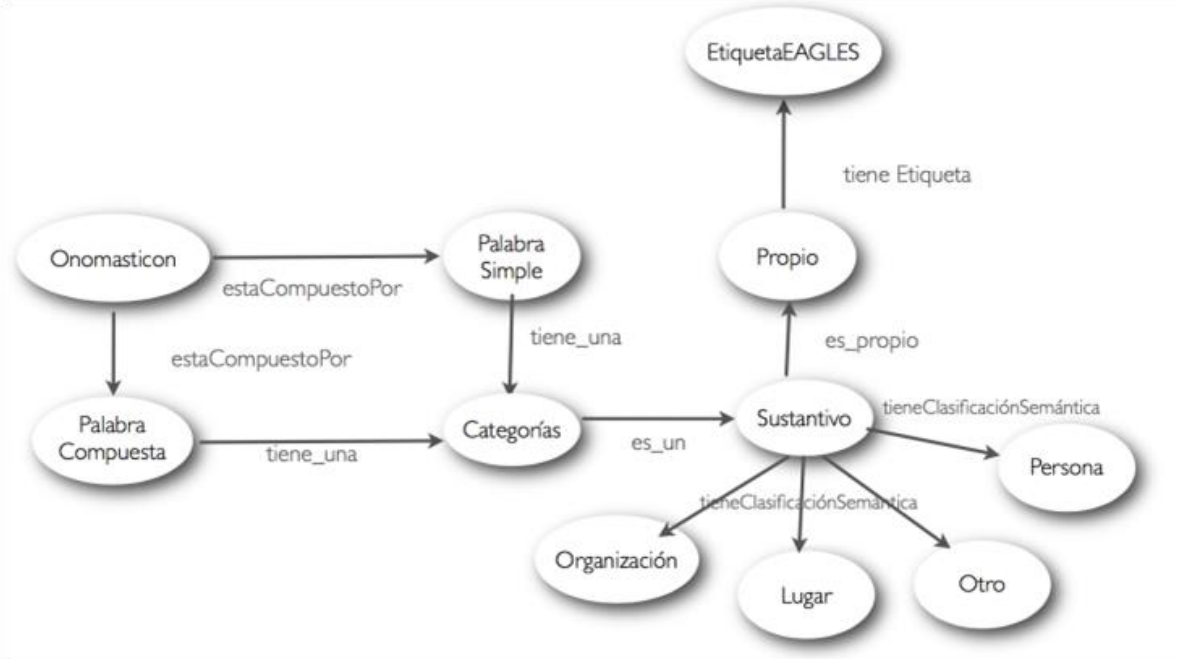


Figura. 3.3 Esquema Conceptual del Onomasticon con las relaciones semánticas

3.2.2. Ontología Lingüística

La ontología lingüística es la encargada de especificar la gramática del lenguaje. Su función es apoyar el proceso de procesamiento del lenguaje natural. Para ello, ella describe de manera general a las unidades léxicas[51] como objetos lingüísticos en una base de datos léxica, y las relaciones entre ellas en una jerarquía conceptual (taxonomía ontológica). Por lo tanto, se puede decir que la ontología lingüística es una representación de conceptos lingüísticos y sus relaciones en un dominio específico, en este caso, el lenguaje español. La ontología lingüística utiliza OWL para las especificaciones de las estructuras sintácticas y semánticas. En general, en la ontología lingüística se formaliza la gramática del español utilizando el enfoque denominado constituyentes[52], como se muestra en la Tabla. 3.1. A continuación describimos cada uno de los componentes que la integra:

[51] Unidades léxicas: que son las unidades mínimas de significación, generalmente más pequeñas que las palabras, como las raíces verbales, las desinencias verbales para la persona, el tiempo, el número, etc.[40]

[52] "es el enfoque donde las oraciones se analizan mediante un proceso de segmentación y clasificación. Se segmenta la oración en sus partes constituyentes, se clasifican estas partes como categorías gramaticales, después se repite el proceso para cada parte dividiéndola en subconstituyentes, y así sucesivamente, hasta que las partes sean palabras indivisibles dentro de la gramática (morfemas)" [12].

- ✓ Conceptos
  - Categorías: los conceptos se agrupan en categorías conocidas: sustantivos, adverbios, adjetivos y verbos
  - Sustantivos: son los pronombres y nombres que denotan cosas.
  - Verbos: denotan sucesos o eventos. A su vez, los verbos se clasifican en verbos transitivos, verbos copulativos y verbos intransitivos.
  - Adjetivos: son los que modifican los sustantivos.
  - Adverbios: son los que modifican el verbo.
  - Frase: las frases se clasifican en Frase Nominal, Frase Adverbial y Frase Verbal.
  - Oración: está compuesta por una Frase Nominal seguida por Frase Verbal.
  - Estructura Sintáctica: cada categoría tiene una estructura sintáctica, el cual se define mediante reglas, por ejemplo: $oracion \rightarrow FraseNominal\ FraseVerbal$
  - Estructura Semántica: cada categoría tiene una estructura semántica

- ✓ Relaciones
  - Tiene: relaciona los conceptos sustantivos, verbos, adverbios y adjetivos con los Conceptos Estructura Sintáctica y Estructura Semántica. A todos esos conceptos se les debe definir tanto su estructura sintáctica como su estructura semántica.
- ✓ Axiomas
  - Los axiomas son un conjunto de reglas que verifica si las sentencias son válidas, tanto sintáctica como semánticamente.

A continuación, se muestran algunos ejemplos de componentes de esta ontología (ver más adelante y en el anexo B otros ejemplos):

"los Verbos_Copulativos tienen una estructura sintáctica como semántica"

$$Verbos\ Copulativos \subseteq Verbos \cap tiene.EstructuraSintactica \cap tiene.EstructuraSemantica$$

Por otro lado, para construir la estructura semántica se usan los atributos usados en la semántica ontológica [40] que se mencionan a continuación:

- ✓ *Agente*: el participante designado por el predicado para hacer o causar la acción. Por lo general el sujeto del verbo transitivo es el agente. Los agentes puede ser las organizaciones, personas, animales, etc.
- ✓ *Paciente*: participante afectado por la acción
- ✓ *Tema*: el participante que cambia de lugar, condición o estado, o que está en determinado estado o posición, es decir el participante manipulado por la acción. Por lo general el tema son los objetos directos de los verbos transitivos, sujetos en las oraciones intransitivas o los complementos verbales.
- ✓ *Instrumento*: El objeto o evento utilizado para llevar a cabo la acción.
- ✓ Experimentador: el participante que se informa de algo o que experimenta un estado psicológico por lo expresado en el predicado

- ✓ *Fuente*: objeto desde el cual ocurre el movimiento
- ✓ *Meta*: objeto hacia el que el movimiento es dirigido.
- ✓ *Lugar*: sitio en que la acción o estado descrito por el predicado tiene lugar.
- ✓ *Beneficiario*: el participante que se beneficia de la acción expresada por el verbo.

Partiendo de lo anterior, un ejemplo de la estructura semántica de los verbos transitivos, para dos de sus casos, es:

1. $EstructuraSemantica \subseteq Agente \cap Tema \cap Beneficiario$
2. $EstructuraSemantica \subseteq Agente \cap Tema$

Y las estructuras sintácticas de 1 y 2, son las siguientes:

1. $EstructuraSintactica \subseteq Sujeto \cap Objecto\_Directo \cap Objecto\_Indirecto$
2. $EstructuraSintactica \subseteq Sujeto \cap Objecto\_Directo$

A continuación, se presentan algunos ejemplos de estructuras sintácticas y semánticas para varios tipos de verbos transitivos.

Tabla 3.1. Ejemplos de la estructura sintáctica y semántica de los verbos transitivos

Verbo	Estructura semántica	Estructura sintáctica
dar	$Agente \cap Tema \cap Beneficiario$ Ejemplo: Juan(Agente) le *dio* las notas(Tema) a los estudiantes(Beneficiario)	$Sujeto \cap Objecto\_Directo \cap Objecto\_Indirecto$ Ejemplo: Juan(Sujeto) le *dio* las notas(Objecto Directo) a los estudiantes (Objeto Indirecto)
comer[53]	$Agente \cap Paciente$ Ejemplo: Pedro(Agente) *comió* galletas(Paciente)	$Sujeto \cap Objecto\_Directo$ Ejemplo: Pedro(Sujeto) c*omió* galletas(Objeto directo)
encontrar	Como localización $Agente \cap Tema \cap Lugar$ Ejemplo: Juan(Agente) encontró a Pedro(Tema) en Mérida(Lugar)	Como localización $Sujeto \cap Objecto\_Directo \cap Objecto\_Indirecto$ Ejemplo: Juan(Sujeto) encontró a Pedro(Objeto directo) en Mérida(Objeto Indirecto)

Como vemos, de esta ontología se derivan un conjunto de reglas para verificar si las consultas son válidas, tanto sintáctica como semánticamente (ver parte de ellas en la Tabla. 3.2). En la Figura. 3.4 se muestra parte de la ontología lingüística en Protégé, donde se observa la jerarquía de conceptos y algunas de las reglas.

[53] El verbo comer puede ser transitivo o intransitivo. En el caso Pedro comió, se comporta como verbo intransitivo y tiene otra estructura sintáctica y semántica

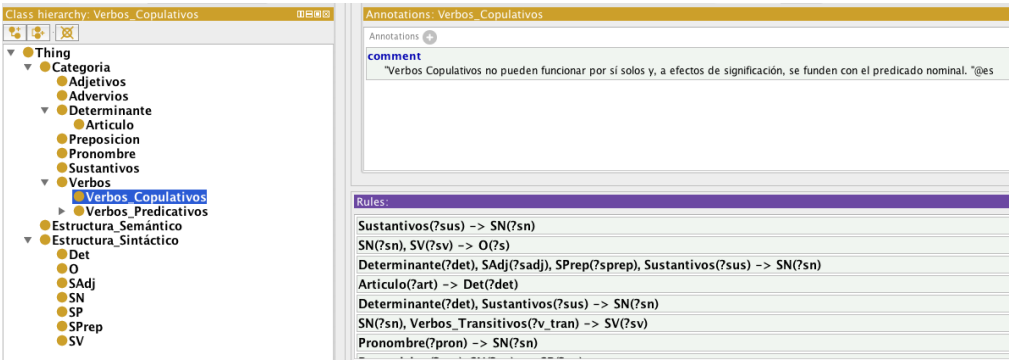


Figura. 3.4 Ontología lingüística en Protégé

Tabla 3.2. Parte de la Gramática usada en la ontología lingüística

Constituyente	Reglas
Oración -> O	O -> (SN) SV
Sintagma Nominal -> SN	SN -> { (Det) N (SP)} | Pron | Art N | Sno
Sintagma Verbal -> SV	SP-> P SN
Sintagma Preposicional -> SP	Sno -> Np
Sintagma Nombre-> SNo	SV -> V (SN) | V SP | V SN SP
Determinante -> Det	V->Vi | Vt | Vc
Nombre -> N	N->Nc | Np
Nombre Común -> Nc	Subj->Np
Nombre Propio -> Np	Ag->Subj
Pronombre -> Pron	
Preposición -> P	
Verbo intransitivo ->Vi	
Verbo transitivo ->Vt	
Verbo copulativo -> Vc	
Articulo->Ar	
Sujeto -> Subj	
Agente->Ag	

Recordemos que, en expresiones regulares, los constituyentes opcionales se representan entre paréntesis -> (), los constituyentes alternativos se separan mediante una -> |, y los grupos de constituyentes se representa -> {}.

2.1.1. Ontología Interpretativa

La ontología interpretativa se diseñó tomando como base la ontología del proyecto mikrokosmo [40] para describir el contexto, y a [1] para describir el perfil del usuario. Para describir el contexto, la ontología Interpretativa está compuesta por:

a) Conceptos

- Entidades: representa objetos físicos como abstractos (normalmente son los sustantivos, adjetivos y adverbios)
- Eventos: representa una acción (normalmente son verbos)

- Relaciones: indican las diferentes relaciones que pueden existir entre los conceptos definidos previamente, o propiedades que los mismos pueden tener.

La clasificación semántica de las Entidades y de los Eventos en el MODS se muestra en las tablas 3.3 y 3.4, respectivamente:

Tabla 3.3. Clasificación Semántica de las Entidades

ENTIDADES del MODS	**Definición [29]**
Abstractos	Los abstractos pueden ser números, conjuntos, definiciones
Concretos	Los concretos son objetos físicos o que se pueden definir en algo específico (por ejemplo, el planeta Venus, ese árbol, aquella persona, organizaciones, etc.)

Tabla 3.4. Semántica de los Eventos

EVENTOS del MODS
Comunicación • General (Decir, hablar) • Valoración (Criticar, felicitar) • Mandato (Suplicar, ordenar)
Comportamiento • General (Comportar) • Relaciones Sociales (Acoger, casar) • Fisiológica (Llorar, Orinar) • Vida (Hacer, Matar)
Cambio • Destrucción Consumo (comer, Eliminar, Gastar) • Cuidado Personal (Lavar, Cepillar) • General (Cocinar, Pintar) • Modificación (Secar, Hervir, Romper) • Creación (Crear, fabricar)

Mental
• Percepción (Ver, escuchar)
• Sentimientos (Sentir)
• Sensación (Gustar, querer, doler)
• Cognición
  ○ General (Entender, Pensar)
  ○ Creencia (Opinar, Creer)
  ○ Conocimiento (Recordar, saber)

Posesión

- General
  - Adeudar
  - Deber
- Pertenencia
  - Poseer
  - Tener
  - Carecer
- Transferencia
  - Recibir
  - Pagar
  - Dar
  - Cobrar

El concepto relación puede ser de dos tipos: relaciones para vincular entidades, entidades y eventos; o propiedades para describir atributos de los conceptos. Algunos ejemplos de las relaciones que vinculan a entidades y eventos son mostrados en la Tabla. 3.5:

Tabla 3.5 Tipos de Relaciones para Vincular entidades y eventos

MODS Relación	**Definición**
Relación_binaria	Donde solo se relaciona dos conceptos. Por ejemplo. Pasajero compró boleto
EsParteDe	Son las relaciones de las partes que componen un todo. Por ejemplo el concepto apéndice es parte del concepto intestino.
Es_Sinónimo_de	Es la relación entre dos términos de significados (conceptos) similares e intercambiables en el discurso por pertenecer a la misma categoría sintáctica. Ejemplo: amplio/extenso, pelo/cabello, estimar/apreciar
Es_Polisemia_de	Es la relación entre un término que tiene varios significados, adquiridos por ampliación o restricción de su significado original, de modo que todos ellos están emparentados semánticamente.

b) Relaciones

Las relaciones entre los conceptos de esta ontología son del tipo “tiene”. Por ejemplo, una entidad *tiene* relaciones con otras entidades, con eventos, una entidad *tiene* propiedades, etc.

La Figura. 3.5. presenta parte de la ontología interpretativa con estos componentes. Además, esta ontología guarda algunas instancias de los objetos definidos por ella, que se van conociendo por el uso del MODS (nombre de organizaciones, sitios, profesiones, etc., ver en la Figura. 3.6 un ejemplo de instancias para el caso de las entidades), que van caracterizando al contexto. Finalmente, la ontología interpretativa es completada con el perfil del usuario, como se muestra en la Figura. 3.7.

Figura 3.5. Ejemplo de Ontología Interpretativa

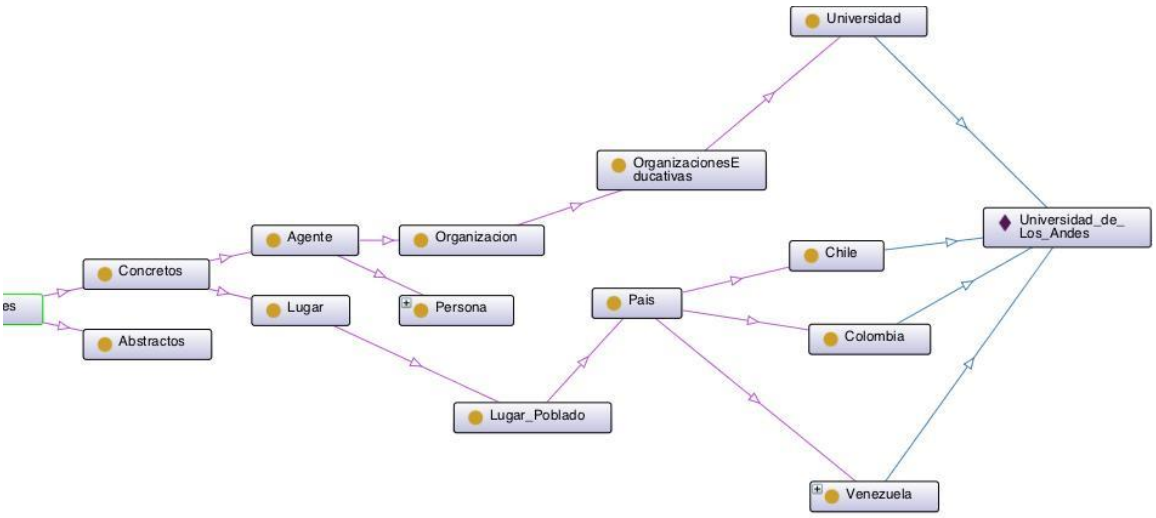


Figura 3.6. Ejemplo de instanciación de la Ontología Interpretativa

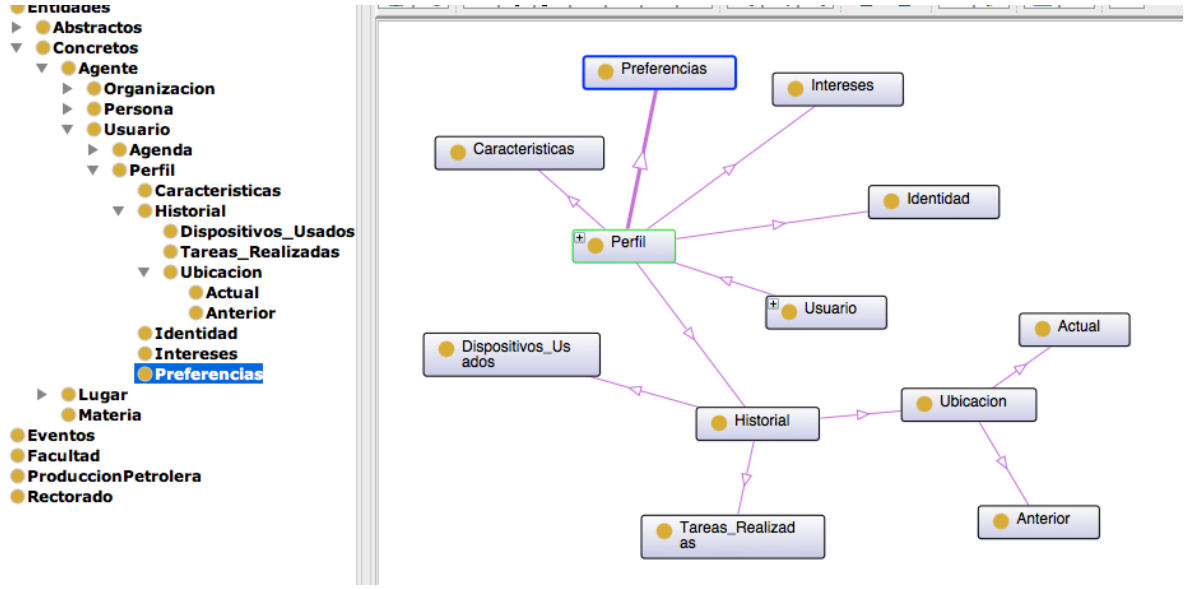


Figura 3.7. Modelo del perfil del usuario en la ontología interpretativa

2.1.2. Ontología de Tareas

En cuanto a la ontología de tarea, se usa OWL como lenguaje de definición de la ontología. Los círculos rojo, azul, verde y morado de la Figura. 3.8 representan el diagrama taxonómico de la ontología de tarea. A continuación, se describen los elementos de esta ontología:

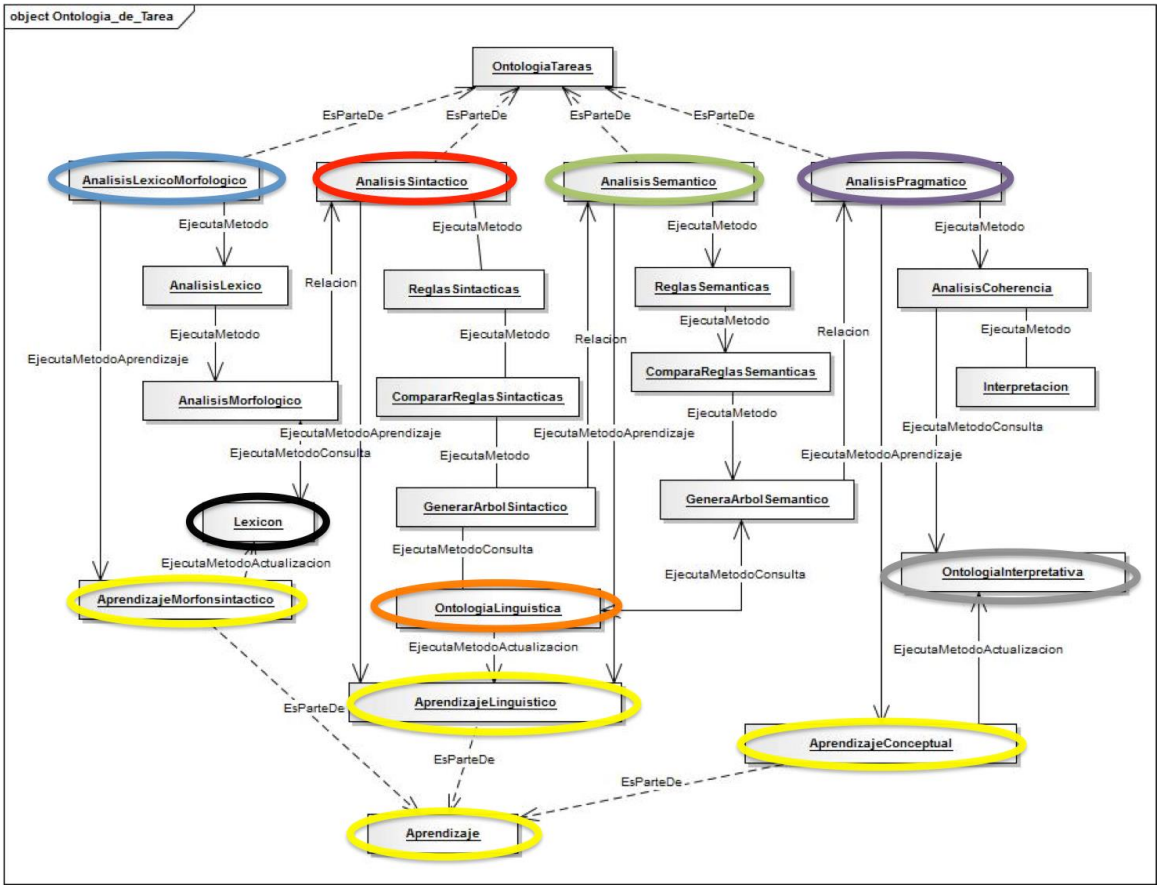


Figura. 3.8. Ontología de Tareas (Integra las otras Ontologías de MODS y del componente de aprendizaje)

Representan las tareas a realizar en cada fase del procesamiento del lenguaje natural:

1. Análisis–léxico-morfológico (concepto del círculo azul): en esta tarea se obtiene la información morfológica contenida en el lexicón (concepto del círculo negro) de cada una de las palabras de la consulta.
2. Análisis-sintáctico (concepto del círculo rojo): en esta tarea se obtiene la estructura sintáctica, utilizando para ello una gramática (almacenada en la ontología lingüística).
3. Análisis-semántico (concepto del círculo verde): en esta tarea se obtiene la estructura semántica (usando las reglas semánticas e información, guardadas en la ontología lingüística e interpretativa).
4. Análisis-pragmático (concepto del círculo morado): en base al contexto y perfil del usuario (definidos en la ontología interpretativa), esta tarea obtiene la interpretación final de la consulta.
5. Los conceptos vinculados a las otras fuentes de conocimiento del MODS (Lexicón (círculo negro), Ontología lingüística (círculo amarillo), Ontología Interpretativa (círculo gris)). Es de esta forma que se da la integración

ontológica, a través de estos conceptos que invocan a esas otras fuentes de conocimiento.

6. Existe un conjunto de conceptos que ayudan al procesamiento del lenguaje natural, que son los que se encargan de obtener las palabras, las morfologías, las reglas sintácticas y semánticas, los árboles de traducción sintácticas y semánticas, los análisis de coherencia y la interpretación de la consulta (estos conceptos no están dentro de un circulo).
7. Finalmente, están los conceptos vinculados al componente de aprendizaje (conceptos con color amarillo), que identifican las diferentes formas de aprendizaje que se pueden realizar en el MODS.

a) Relaciones:

Las relaciones son las siguientes:

Relaciones entre conceptos:

- Es_parte_de: relación que establece los componentes de la ontología de tareas.
- EjecutaMetodo: Esta relación es la que invoca los conceptos de las tareas de análisis con los que se realiza el procesamiento.
- EjecutaMetodoConsulta: Esta relación es la que consulta los otros niveles de conocimiento (Lexicón, Ontología lingüística, Ontología Interpretativa) del MODS.
- EjecutaMetodoAprendizaje: Esta relación invoca el proceso de aprendizaje cuando EjecutaMetodoConsulta indica que no conoce el término, o la estructura sintáctica, o la estructura semántica, o el concepto.
- EjecutaMetodoActualizacion: Esta relación actualiza cualquier nivel de conocimiento (ontología lingüística, ontología interpretativa, lexicón) del MODS.

Propiedades de los objetos:

1. ObjetivoMetodo: Esta propiedad es la descripción de cada método utilizado las tareas de análisis.
2. EntradasMetodos: Son las entradas necesarias para la ejecución del método invocado.
3. SalidasMetodos: Son las salidas del método en ejecución.
4. Axiomas

La representación de la Ontología de Tareas se realizó con lógica descriptiva, a continuación se muestran un ejemplo de los axiomas en su base de conocimiento (ver en el anexo C el resto de los axiomas).

$$AnalisisLexicoMorfologico \subseteq Tarea \cap \forall EjecutaMetodo.AnalisisLexico \cap \forall EjecutaMetodoAnalisisMorfologico$$

Este axioma indica que para que se ejecute la tarea de AnalisisLexicoMorfologico se deben ejecutar los métodos de Análisis Léxico y Morfológico.

3.3. Descripción detallada del MODS

Como se puede observar en la Figura. 3.8, la ontología de tareas es la ontología que se encarga de integrar los otros tres niveles ontológicos que presenta la arquitectura (lexicón, lingüística e interpretativa). A continuación, se explica el proceso de integración en forma general, y su relación con los otros componentes del MODS:

a) La primera tarea a realizar es el análisis léxico-morfológico, y se usa como componente de apoyo al lexicón (ver círculo color azul para el análisis léxico-morfológico y color negro para el lexicón en la Figura. 3.8). En esta tarea se obtiene la información morfológica de cada una de las palabras que se encuentran en la consulta, y con el onomasticon se determina si hay nombres propios en la consulta. El análisis léxico morfológico se divide en las siguientes sub-tareas: i) Análisis léxico, su principal función es de recorrer la consulta de entrada y separarla en componentes léxicos (tokens en inglés); ii) análisis morfológico, el cual se apoya en el lexicón, y consiste en determinar la forma, clase o categoría gramatical de cada palabra de la consulta.

b) El siguiente paso es el análisis sintáctico (ver circulo de color rojo en la Figura 3.8), el cual, a partir de la salida del proceso anterior, y con el fin de detectar oraciones o frases significativas para el lenguaje (en nuestro caso español), usa una gramática. El proceso de “traducción” compara las reglas que se encuentra en la ontología lingüística (ver circulo color anaranjado en la Figura 3.8) contra las palabras del texto de entrada, cada regla que es instanciada por las palabras del texto agrega un elemento a la estructura o la termina de generar. La estructura que se produce como resultado de estas instanciaciones de las reglas es el “árbol de traducción”

c) Después, con la ontologías lingüística e interpretativa (ver círculos naranja y gris en la Figura 3.8) se realiza el análisis semántico (ver circulo verde en la Figura 3.8), el cual es el encargado de establecer qué combinaciones de significados de palabras individuales son posibles a la hora de crear un significado coherente en una oración, lo que puede reducir el número de posibles significados para cada palabra en una oración determinada. De esta forma, el analizador obtendrá la esencia de la pregunta. El análisis semántico representa uno de los elementos claves del “conocimiento” del MODS, y en función a su variedad y detalle será la riqueza del vocabulario, expresión, entendimiento, respuesta y utilidad que el propio análisis ofrezca.

d) La última tarea de la ontología de tareas es el análisis pragmático o del contexto (ver círculo morado en la Figura 3.8). Dado que la semántica se ocupa del significado literal de las expresiones lingüísticas, la pragmática trata el significado adicional que adquieren las consultas en el contexto del usuario. Por lo

tanto, el analizador pragmático utiliza la estructura semántica obtenida en la tarea anterior para desarrollar la interpretación final de la consulta, en función del contexto. En este nivel se analizan los mecanismos de coherencia del discurso del usuario (es decir, los elementos lingüísticos que utiliza para comunicarse en la Web), cuál es su intención comunicativa, etc., usando para ello la meta-ontología interpretativa (ver círculo color gris en la Figura 3.8). En esta fase se cubren aspectos tales como la identificación de objetos referenciados por determinados constituyentes de la frase (sintagmas nominales, pronombres, elementos elididos, etc.), el análisis de aspectos temporales, la identificación de la intención del usuario (temas y focos), para interpretar apropiadamente la consulta dentro del dominio específico vinculado al usuario. Así, el análisis pragmático realiza la interpretación final de la consulta en función del contexto y del perfil del usuario.

Otros componentes de la Figura 3.8 tienen que ver con los vinculados al proceso de aprendizaje del MODS (círculos amarillos), que permiten mantener actualizado a todas las ontologías. Este componente será explicado en detalle en el siguiente capítulo.

3.4. Proceso de Interpretación de la Consulta

Para entender el proceso de interpretación de la consulta realizado por la ontología de tarea, se realiza un ejemplo de una consulta sencilla, por ejemplo “Universidad de Los Andes”.

Como se mencionó en la sección anterior, el proceso comienza con una consulta en lenguaje natural, realizando la primera fase del análisis léxico morfológico, utilizando los recursos lingüísticos de las ontologías de Lexicón y Onomasticon, generando el grafo de la consulta mostrada en la Figura 3.9.

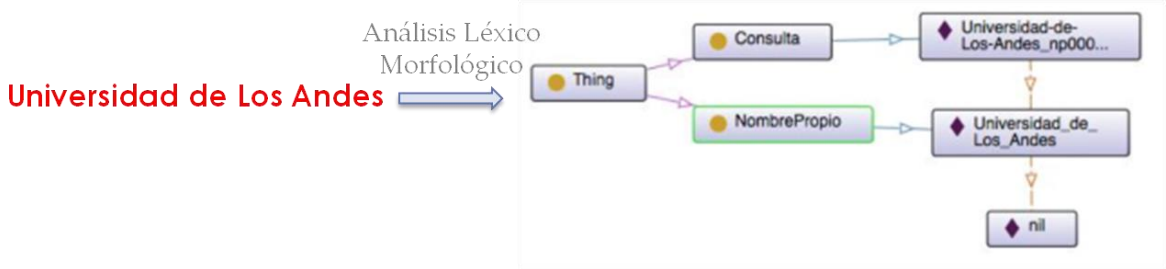


Figura. 3.9. Grafo generado después de realizar el Análisis Léxico Morfológico

Como se puede observar en la Figura 3.9, la consulta es etiquetada usando la etiquetación léxica EAGLES, y, además, es separada en tokens, donde el individuo nil es el fin de la consulta. También, indica que Universidad de Los Andes es un Nombre Propio. Siguiendo con el proceso, se realiza el Análisis Sintáctico, donde se usa como recurso la ontología lingüística; este proceso recibe como entrada la salida del proceso de Análisis Léxico Morfológico, y genera el grafo de la consulta mostrada en la Figura 3.10.

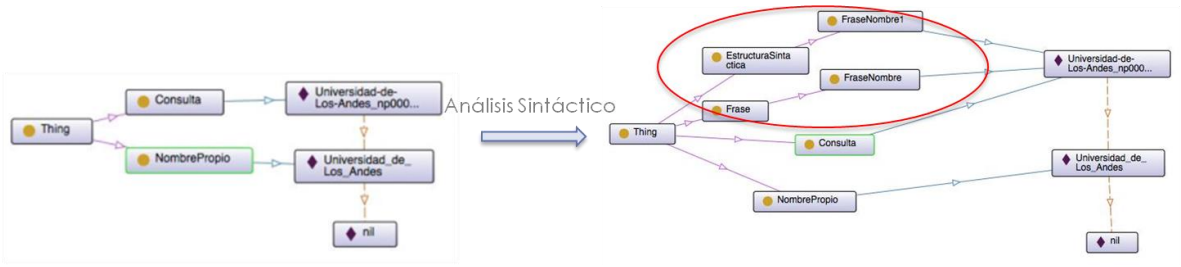


Figura. 3.10. Grafo generado después de realizar el Análisis Sintáctico

En la Figura 3.10, lo que está en círculo rojo refleja la nueva información en el grafo de consulta, que consiste en la estructura sintáctica de la consulta y qué tipo de frase es, después del razonamiento de las reglas que están definidas en la ontología lingüística. Siguiendo con el proceso, se realiza el Análisis Semántico, recibiendo como entrada el resultado obtenido del Análisis Sintáctico y utilizando la ontología lingüística. Como resultado de ese proceso, se obtiene el grafo de la consulta mostrado en la Figura 3.11.

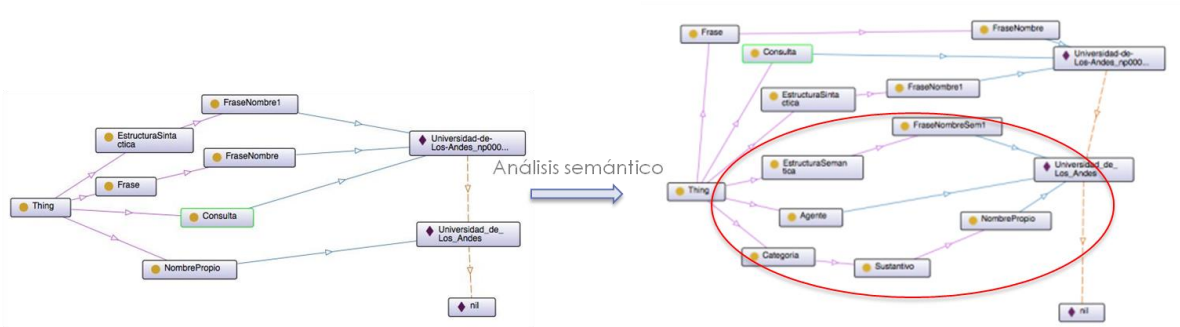


Figura. 3.11. Grafo generado después de realizar el Análisis Semántico

En la Figura 3.11 se puede observar lo que está en el círculo rojo, que representa la nueva información, que consiste en la estructura semántica, la cual indica que el termino Universidad de Los Andes está actuando como agente según su rol semántico, y también define la categoría del término. Todo esto se obtiene después de haber hecho el proceso de razonamiento en la ontología lingüística. Finalmente se realiza el proceso del Análisis Pragmático, que recibe como entrada el grafo generado del Análisis Semántico y usa la ontología interpretativa, obteniendo el resultado de la Figura 3.12.

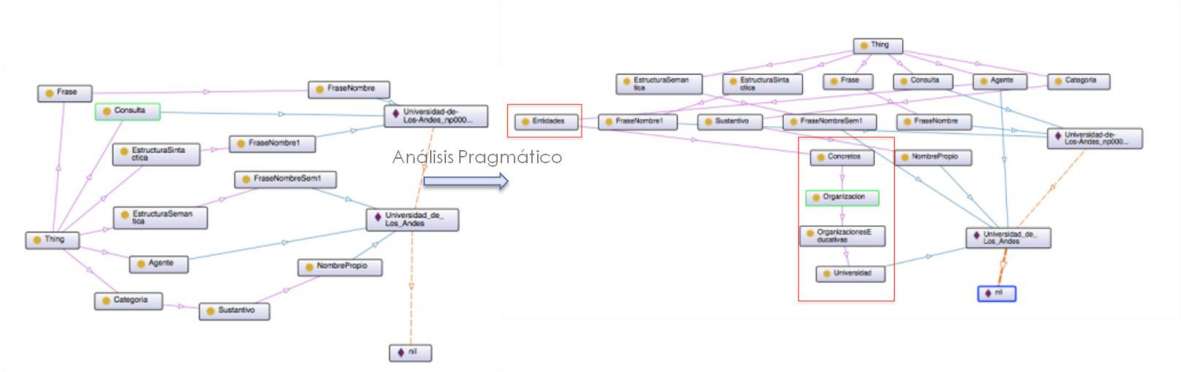


Figura. 3.12. Grafo generado después de realizar el Análisis Pragmático

En la Figura 3.12 se muestra el resultado de la interpretación de la consulta, donde nos dice que el Termino Universidad de Los Andes es una Universidad, Universidad es una Organización Educativa, etc. La información nueva esta en los rectángulos rojos.

El resultado de la consulta está interpretado en un lenguaje ontológico, que es el objetivo del MODS. Una vez que la consulta está interpretada se pasa por un proceso de transformación, en nuestro caso es una consulta Booleana, para ser enviada a la Web usando el motor de búsqueda seleccionado para tal fin, y así el usuario final obtiene la información que desea buscar en la Web. Los resultados que genera la Web pasan por el componente de aprendizaje, que es un proceso interno del MODS, para lograr la adaptabilidad de los componentes del MODS, este proceso se explica en el capítulo 4.

Capítulo 4: Aprendizaje Ontológico

Para lograr el proceso de adaptación del MODS se requiere de un componente de aprendizaje ontológico que realice el proceso de adquisición de conocimiento, y posteriormente la actualización de los mismos. Sin este componente de aprendizaje, el MODS se convierte en un simple sistema de interpretación de consultas, usando un conocimiento inicialmente definido en sus componentes. Este capítulo presenta el componente de aprendizaje del MODS, para lo cual se organiza de la siguiente manera: en la sección II se presentan los antecedentes en cuanto a aprendizaje de ontologías, en la sección III se presenta el modelo general del componente de aprendizaje, la sección IV describe el proceso de aprendizaje morfosintáctico, y finalmente, la sección V presenta el proceso de aprendizaje semántico.

4.1. Antecedentes

El MODS tiene dos tipos de aprendizajes, los cuales se mencionan a continuación:

- ✓ *Aprendizaje de Información Morfosintáctica*: consiste en una serie de métodos de aprendizaje para la adquisición, ampliación y refinamiento de la información morfológica y sintáctica almacenada en el MODSLex[54], sin los cuales MODS no podría robustecerse en su capacidad de interpretación del lenguaje natural. Las estructuras del lenguaje natural principales que tienen que ver con el aprendizaje morfosintáctico son las unidades léxicas de alto nivel (verbos, sustantivos, adjetivos, adverbios).
- ✓ *Aprendizaje de Información Semántica*: Este aprendizaje cubre el aprendizaje de conceptos, relaciones taxonómicas, relaciones no taxonómicas, propiedades y axiomas. En específico, de especial interés para el MODS es el aprendizaje de relaciones no taxonómicas, que consiste en el descubrimiento de verbos relacionados con el dominio, la extracción de conceptos que están relacionados no taxonómicamente, de reglas de comportamiento, entre otras cosas. De la misma manera que el aprendizaje de información morfosintáctica es clave para el MODSLex, el aprendizaje de información semántica es fundamental para las ontologías interpretativa (dominio) y lingüística.

El nuevo conocimiento descubierto se incorporará en las estructuras del MODS respectivas, como se muestra en la Tabla. 4.1.

[54] El lexicón del MODS

Tabla 4.1. Conocimiento Descubierto y Componente MODS donde es incorporado.

Conocimiento Descubierto	Proceso de aprendizaje que lo descubre	Componente MODS que impacta
Información Léxica	Aprendizaje Morfosintáctico	Lexicón
Información de Dominio	Aprendizaje Semántico	Ontología Interpretativa
Información lingüística	Aprendizaje Semántico	Ontología Lingüística

Los métodos/técnicas de aprendizaje/descubrimiento de conocimiento para el MODS son muy diversos, y son definidos en función del objeto de aprendizaje y la fuente de información a usar.

Los trabajos más relevantes para nuestra investigación relacionados a los problemas de aprendizaje morfosintáctico y semántico, se mencionan brevemente a continuación:

4.1.1. Aprendizaje Morfosintáctico

- ✓ DAILLE [20]: Su objetivo es la extracción de términos. La principal idea de este sistema es combinar conocimiento lingüístico con cálculos estadísticos. Se inicia con un corpus que contiene toda la información morfológica. Luego se genera una lista de términos candidatos de acuerdo con unos patrones de formación de términos. A continuación se usan métodos estadísticos para filtrar esa lista y quedarse con los términos aprendidos.
- ✓ TERMS [21]: Se basa en la idea que los términos suelen repetirse en documentos técnicos con más frecuencia que sintagmas nominales no terminológicos, y que los sintagmas nominales que representan términos tienen una estructura diferente que los sintagmas nominales que no los representan. Normalmente, los sintagmas que representan términos no suelen incluir modificadores ni determinantes, al contrario que los sintagmas nominales comunes. Así, TERMS utiliza un filtro estadístico que rechaza las cadenas que aparecen sólo una vez en un documento.

4.1.2. Aprendizaje Semántico

- ✓ ONTOLEARN [35]: Es un sistema que permite enriquecer una ontología de dominio con conceptos y relaciones. Usa técnicas de aprendizaje de máquina (algunas de las técnicas usadas para el descubrimiento de conocimiento (patrones) son las cadenas de markov[55]) y procesamiento del lenguaje natural para este objetivo.

[55] Una cadena de Márkov es una serie de eventos, en la cual la probabilidad de que ocurra un evento depende del evento inmediato anterior. En efecto, las cadenas tienen memoria: "recuerdan" el último evento y esto condiciona las posibilidades de los eventos futuros.

- ✓ DL-LEARNER [23]: Es un sistema que aprende una definición de un concepto en lógica descriptiva[56]. Este sistema se basa en la programación lógica inductiva (PLI). El sistema recibe como entrada una base de conocimiento y un conjunto de entrenamiento con ejemplos positivos y negativos de instancias del concepto de interés. A partir de allí realiza un proceso de aprendizaje inductivo[57] del concepto.
- ✓ SVETLAN [10]: Es un sistema para construir una ontología a partir de jerarquías de nombres. Toma dominios semánticos como unidades temáticas, y construye dominios estructurados que clasifican los nombres en esos dominios de acuerdo con los verbos y las relaciones entre ellos.
- ✓ HASTI [47]: Es un sistema para la construcción automática de ontologías, usa como entrada información no-estructurada en la forma de texto en lenguaje natural persa. HASTI no usa conocimiento previo, es decir, construye las ontologías desde cero. Usa un lexicón que se encuentra inicialmente vació, y va creciendo incrementalmente en la medida que va aprendiendo nuevas palabras. HASTI aprende conceptos, relaciones conceptuales taxonómicas, no-taxonómicas, y axiomas, con lo cual va construyendo la ontología sobre un núcleo/kernel inicial. HASTI usa un enfoque simbólico hibrido (una combinación de lógica, lingüística, y métodos heurísticos, entre otros).
- ✓ TEXT-TO-ONTO [25]: Su principal objetivo consiste en el aprendizaje de relaciones taxonómicas y no-taxonómicas. Se basa en análisis estadístico y un analizador/parsing sintáctico de textos en francés, más reglas de asociación y de poda/acotación. Aprende desde texto libre, semi-estructurado, desde otras ontologías y bases de datos. El resultado del proceso de aprendizaje es una ontología de dominio que contiene conceptos específicos del dominio bajo estudio. Todo el proceso es supervisado por el ingeniero ontológico.
- ✓ MARSID [34]: Es un sistema para el descubrimiento de reglas de asociación entre un conjunto de Ítems en una gran base de datos. Dada una base de datos de las transacciones de los clientes de un almacén, donde cada transacción consiste en la canasta de compra de un cliente en una visita, en este trabajo se plantea un algoritmo que genera todas las reglas de asociación significativas entre los artículos en la base de datos. El algoritmo incorpora técnicas de estimación y de poda.

4.2. Arquitectura del Componente de Aprendizaje Ontológico del MODS

El aprendizaje ontológico es usado para la adaptabilidad del MODS. Este componente recibe como entrada términos desconocidos o documentos Web (ver Figura. 4.1). El primer caso ocurre con el análisis de la consulta, cuando se

[56] La lógica descriptiva (Descripción Logic:DL) es un formalismo de representación de conocimiento, muy usado en los lenguajes ontológicos actuales.

[57] El aprendizaje inductivo consiste en que el aprendiz realiza un proceso que parte de la observación y el análisis de una característica de lo que observa (por ejemplo, la lengua), hasta la formulación de una regla que la explique. De esta manera, el proceso de aprendizaje va de lo concreto (lo que se observa) a lo general y abstracto (las reglas de lo observado)..

encuentra frente a un término desconocido por el lexicón del MODS (MODSLex); en este caso, el MODS invoca el *aprendizaje ontológico morfosintáctico*. El segundo caso ocurre cuando el proceso de interpretación de la consulta es exitoso, lo que resulta en un grupo de información recuperada de la Web por la consulta generada por MODS. En este caso, el MOD invoca el *aprendizaje ontológico semántico*, el cual aprende nuevos conceptos, relaciones taxonómicas y no taxonómicas.

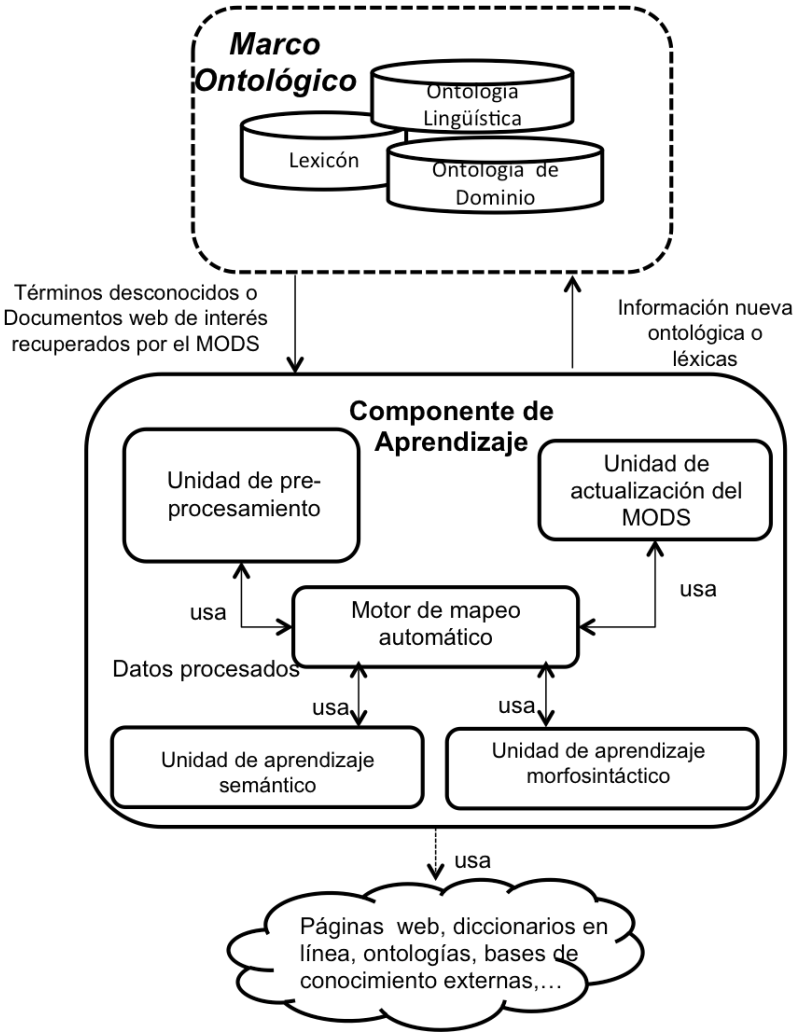


Figura. 4.1. Componente de Aprendizaje del MODS.

La arquitectura del componente de aprendizaje ontológico consta de cinco componentes: el primer componente es la *unidad de pre-procesamiento*, que caracteriza la información de entrada, sea un término desconocido o información recuperada de la Web. El segundo componente es la *unidad de aprendizaje semántico*, el cual contiene un repositorio de técnicas de aprendizaje de conceptos, relaciones taxonómicas y no taxonómicas, a partir de documentos. El tercer componente es la *unidad de aprendizaje morfosintáctica*, este componente utiliza diccionarios en línea para determinar las estructuras léxicas o sintácticas de

términos desconocidos. El cuarto componente es la *unidad de actualización,* que determina el elemento del MODS que se va actualizar según el aprendizaje que se esté realizando: al lexicón, a la ontología interpretativa, o a la ontología lingüística. Finalmente, el *motor de mapeo automático* es el que se encarga de llamar a las unidades de aprendizaje respectivas según la entrada al sistema (determina cuál de los dos aprendizajes se va utilizar).

A continuación se describen las dos unidades de Aprendizaje, que definen los dos tipos de aprendizajes de MODS.

4.3. Unidad de Aprendizaje Morfosintáctico

El componente de aprendizaje morfosintáctico tiene por objetivo enriquecer el lexicón definido para el MODS. En cuanto al aprendizaje morfosintáctico [42], el MODS requiere de un conjunto de conocimientos según la categoría lingüística a aprender. Dicha información es caracterizada en la siguiente función, que constituye la interfaz de MODS con el componente de aprendizaje:

Lex_morf[58](componente léxico, categoría, tipo, genero, numero, modo, tiempo, aspecto, voz, persona)

Donde:

- ✓ Componente léxico: es una de las palabras de la consulta.
- ✓ Categoría: es el tipo de palabra (nombre, verbo, artículo, etc.)
- ✓ Género: indica si pertenece al masculino o al femenino.
- ✓ Número: indica si el objeto nombrado es uno o más de uno. En específico, en español hay dos tipos: singular y plural.
- ✓ Modo: se refiere a la actitud del hablante ante lo que dice. Hay tres modos verbales: Indicativo (por ejemplo. yo he llegado a la ciudad). Subjuntivo (por ejemplo, quizás llegue a la ciudad). Imperativo (por ejemplo, ven aquí)
- ✓ Tiempo: Es la capacidad que tiene el verbo para situar la acción en un contexto temporal determinado. Normalmente, un verbo expresa nociones que se sitúan en el presente, en el pasado o en el futuro.
- ✓ Aspecto: expresa si la acción del verbo ha acabado o tiene sentido durativo.
- ✓ Voz: indica si el sujeto realiza la acción (sujeto activo o agente), o sufre la acción que realiza otro (sujeto pasivo o paciente). En el primer caso decimos que el verbo está en voz activa; cuando el sujeto es paciente el verbo está en voz pasiva
- ✓ Persona: Primera, segunda y tercera.

Por ejemplo, para un nombre o sustantivo desconocido como escuela, la entrada al componente de aprendizaje seria:

[58] lex_mor es el nombre definido para la estructura definida para el lexicón del MODS

lex_morf('escuela', 'desconocido', NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL)

para el cual se esperaría que el componente de aprendizaje devuelva la información mostrada en la Figura. 4.2 (un grafo RDF con la información léxica-morfológica de la palabra escuela).

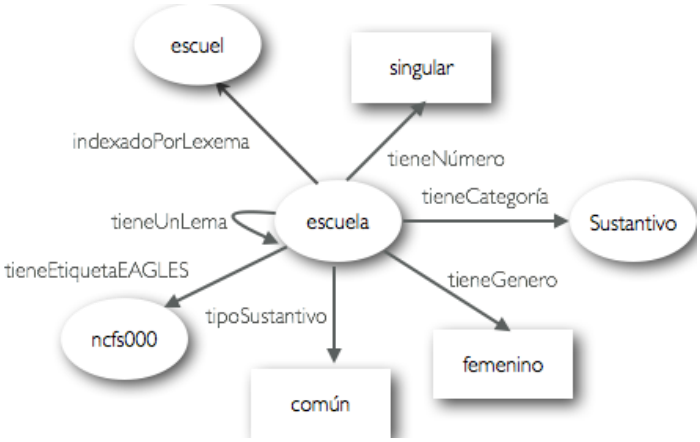


Figura 4.2 Grafo generado por el aprendizaje morfosintáctico

En general, la unidad de aprendizaje morfosintáctico recibe una cadena que contiene una palabra desconocida, y proporciona una de las siguientes salidas:

1. Un grafo RDF de aprendizaje lex_mor con la nueva información léxica descubierta (ver Figura. 4.2).
2. Un mensaje de error, que puede venir acompañado por un archivo con evidencias de uso en la Web de palabras muy semejantes a la buscada.

El componente de aprendizaje morfosintáctico lo constituyen cuatro módulos (ver Figura. 4.3): un primer módulo de pre-procesamiento y caracterización, denominado *aprendizaje simiente* (por proporcionar la semilla para el aprendizaje), un segundo módulo llamado *aprendizaje ágil* (por consultar información rápidamente directamente desde la Web), un tercer módulo llamado *aprendizaje duro* (por requerir procesamiento del lenguaje natural), y un último *módulo de enriquecimiento*, refinamiento o actualización de MODS. A continuación pasamos a detallar cada módulo:

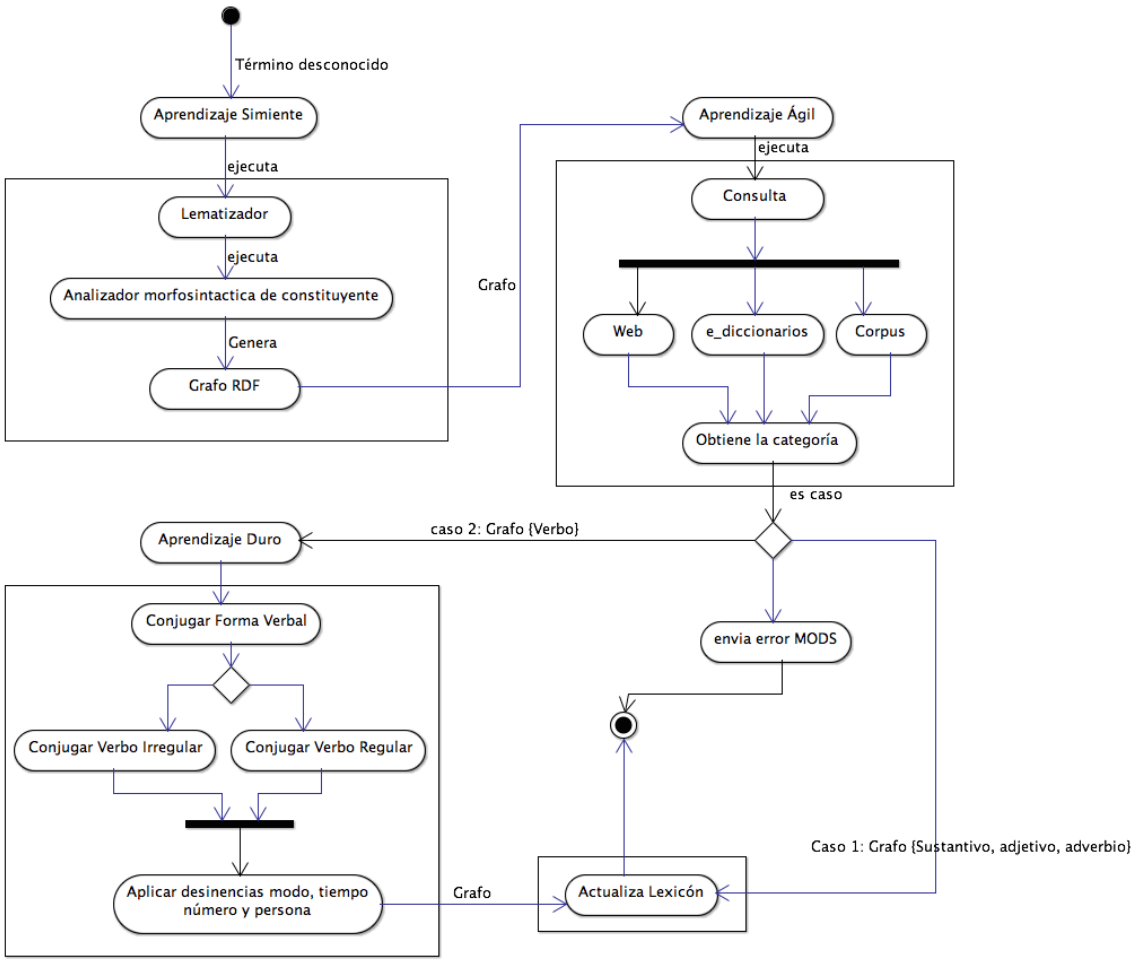


Figura 4.3. Arquitectura de la Unidad de Aprendizaje Morfosintactico del MODS.

El *módulo de aprendizaje simiente* recibe el término desconocido, y halla la información morfosintáctica básica (semilla) del término, tal como su forma canónica, afijos (prefijos y sufijos), etc. Con esta información crea el grafo básico que caracteriza el término (ver Figura. 4.2). La forma canónica de un verbo en infinitivo es el mismo, y para un verbo conjugado su forma infinitiva; para un sustantivo, adjetivo o adverbio en plural, su forma canónica es su singular. El módulo usa una librería de algoritmos diseñados para tales fines: un lematizador [48] y un analizador morfosintáctico [59], respectivamente. En el caso de los adverbios, muchos son derivados de los adjetivos (por ejemplo, felizmente es derivado del adjetivo feliz). Es así que para el aprendizaje simiente de un adverbio se separa el sufijo o el prefijo del núcleo. En cuanto a los adjetivos, son palabras que nombran o indican cualidades, rasgos y propiedades de los nombres o sustantivos a los que acompañan. El tratamiento de aprendizaje simiente para el

[59] https://opennlp.apache.org/

adjetivo consiste en aprender su forma canónica. Las demás unidades léxicas de orden inferior, tales como pronombres, determinantes etc., se encuentran predefinidos en MODS, y no son objeto de aprendizaje.

Después sigue el *módulo de aprendizaje ágil*, cuyo objetivo es validar y aprender información gramatical en función del término. Para ello recibe el grafo generado en el aprendizaje simiente y devuelve alguna de las salidas siguientes (esas salidas son producto de un proceso de enriquecimiento de la entrada recibida, y se lleva a cabo por consultas a diccionarios, corpus, textos, páginas Web, con contenido que se refieren al uso de ese término, entre otras cosas):

- ✓ *Caso 1*: Ocurre cuando la categoría descubierta de la palabra desconocida es un sustantivo, adjetivo o adverbio. En estos tres casos, el grafo de aprendizaje contiene la información gramatical de la palabra, por lo cual no añade más información.
- ✓ *Caso 2*: Ocurre cuando la categoría de la palabra descubierta es un verbo, en ese caso el grafo de aprendizaje es enriquecido con la información gramatical propia del verbo descubierta en las consultas a los diccionarios, corpus, etc.
- ✓ *Caso 3*: Ocurre cuando no se ha encontrado ninguna información relacionada con la palabra desconocida (por ejemplo, cuando se tienen palabras mal escritas, no se tienen referentes de su uso en el español, etc.).

El módulo que continua en el proceso de aprendizaje morfosintáctico es el *aprendizaje duro*. Este módulo está pensado para soportar el aprendizaje morfosintáctico de verbos, y consiste en adquirir nueva información léxica, tal como la conjugación del verbo, entre otras cosas. En general, una de las tareas más complejas dentro del aprendizaje morfosintáctico es el aprendizaje de verbos. En este trabajo se ha desarrollado para el aprendizaje duro un conjunto de algoritmos que pueden diferenciar entre verbos regulares e irregulares, y desplegar la conjugación de los verbos regulares y parte de los irregulares del español, a partir de su forma en infinitivo. En general, el verbo es una categoría léxica que, al conjugarse puede variar en persona, número, tiempo, modo, aspecto y voz. Desde un punto de vista morfológico, el verbo consta de dos partes:

- *La raíz*, que es la parte que nos aporta el significado del verbo, tal y como lo podemos encontrar en el diccionario (información léxica).
- *Las desinencias*, que son los morfemas flexivos que nos dan la información referente a la persona, número, tiempo, modo, aspecto y voz (información gramatical).

De este modo, en una forma verbal como “amo” se pueden distinguir las siguientes partes: am-, raíz cuyo significado es ‘tener amor a alguien o algo’, y la desinencia-que nos revela que la forma verbal amo se corresponde con la primera persona (persona) del singular (número) del presente (tiempo) del indicativo (modo) de la voz activa.

Específicamente, en el caso del verbo regular va directamente al módulo de “conjugar Verbo Regular”, ya que las reglas de conjugación son uniformes; en

cambio, para el caso en que el verbo es irregular, al momento de entrar al módulo de “conjugar Verbo Irregular” deben llamarse otros módulos para poder realizar la conjugación: “catalogar grupo verbal”, "cargar reglas de Irregularidad”, “cargar Patrón de Irregularidad” y “Modificar Raíz”; esto es debido a que cuando el verbo es irregular las reglas de conjugación no son uniformes y varían según el verbo.

En particular, el modulo “conjugar Verbo Regular” consiste en la conjugación de los verbos regulares que su infinitivo termina en ar, er, ir, y según la terminación se obtiene el patrón de conjugación. En cuanto al módulo “conjugar Verbo Irregular”, consiste en la conjugación de los verbos irregulares siguiendo las reglas de irregularidad (de ellas se obtienen, por ejemplo, los cambios ortográficos, los cambios en verbos con raíz que terminan en vocal, etc.), y los patrones de irregularidad (por ejemplo, el patrón To (si la silaba tónica cae en la raíz ar, er, ir, por ejemplo el verbo dormir se cambia la o => ue, duermo, duermes, duerme, duermen, etc.), el patrón Te (cuando la silaba tónica cae en la raíz y la desinencia comienza por e (ir), etc.) [4].

Finalmente, el *módulo de Actualización* del lexicón recibe como entrada la información que describe la palabra desconocida aprendida de los módulos de aprendizaje precedentes en el grafo, e incluye la nueva información léxica en los argumentos vacíos de lex_mor; o construye un nuevo lex_mor, si no existe en el lexicón ese término, con la nueva información descubierta de la palabra desconocida. Finalmente, la función lex-mor actualiza el lexicón.

4.4. Unidad de Aprendizaje Semántico

En la Figura. 4.4 se muestra la arquitectura de la Unidad de Aprendizaje Semántico, el cual ocurre cuando el proceso de interpretación de la consulta es exitoso, generando como resultados un conjunto de enlaces Web, a partir de la consulta hecha por el MODS. Ese conjunto de enlaces es la entrada a esta unidad. La información de esos enlaces puede traer consigo información heterogénea: textos, diccionarios, bases de conocimiento, información semi-estructurada (anotaciones RDF de recursos WEB, esquema XML), información relacional, ontologías, entre otros. Dependiendo del tipo de información de entrada, el proceso semántico de aprendizaje semántico varía.

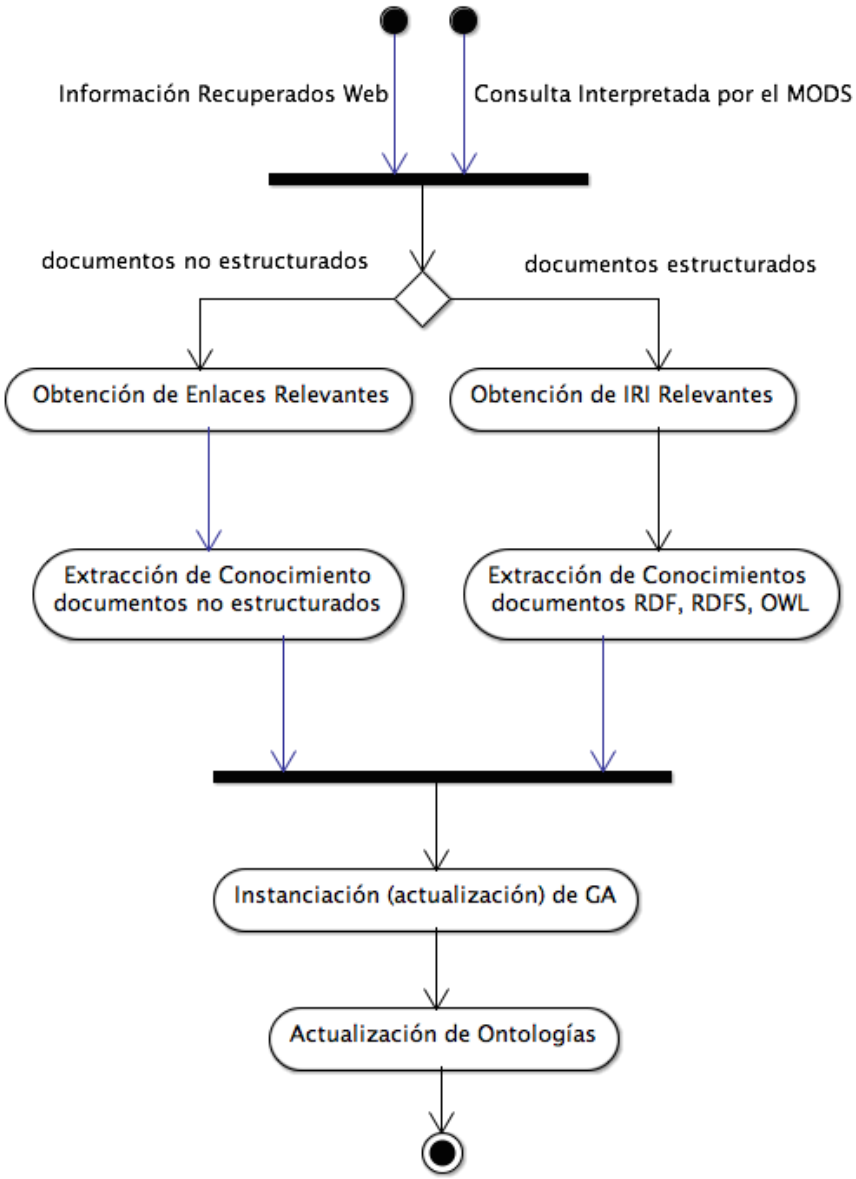


Figura 4.4 Arquitectura para el aprendizaje semántico del componente de MODS.

El lado izquierdo representa el proceso de aprendizaje usando documentos no estructurados[60], y el lado derecho usando documentos estructurados[61]. Los pasos son muy similares, pero los mecanismos usados en cada paso son diferentes. En este trabajo nos centraremos en los documentos no estructurados, por ser estos los documentos que mayormente se recuperan de la Internet. A continuación pasamos a describir los componentes de esa unidad para ese caso (lado izquierdo).

[60] Se trata de los documentos escritos en lenguaje natural, normalmente descritos usando html en Internet.
[61] Se trata de los documentos que tiene una estructura definida, tales como XML, RDF, OWL, etc.

El esquema conceptual de todos los elementos involucrados en el proceso de aprendizaje usando documentos no estructurados que se propone, se muestra en la Figura. 4.5. Partiendo de la hipótesis de que “las entidades son sustantivos y los verbos son eventos en las oraciones que describen el texto” (ver [44]), se define dicho esquema, cuyos componentes se muestran a continuación:

4.4.1. Reconocimiento de los términos en el texto

La interpretación del texto se inicia con el reconocimiento de los términos encontrados en el texto, los cuales pueden ser palabras simples (por ejemplo, universidad, Venezuela) o palabras compuestas (por ejemplo “Universidad de Los Andes”, “San Cristóbal”) (ver figura 4.5). Cada palabra simple tiene un lema (por ejemplo, Universidad tiene como lema Universid). Además, tanto las palabras simples como compuestas tienen una Categoría (representa el tipo de la palabra: si es un verbo, un sustantivo, si el sustantivo es de tipo común (por ejemplo Universidad) o propio (representa los nombres propios, por ejemplo Universidad de Los Andes), si el sustantivo es una entidad, etc. Finalmente, cada una de las entidades tiene una clasificación semántica (por ejemplo, personas, lugares, organizaciones, u otros)). Para este proceso, los recursos lingüísticos utilizados son el lexicón y el Onomasticon.

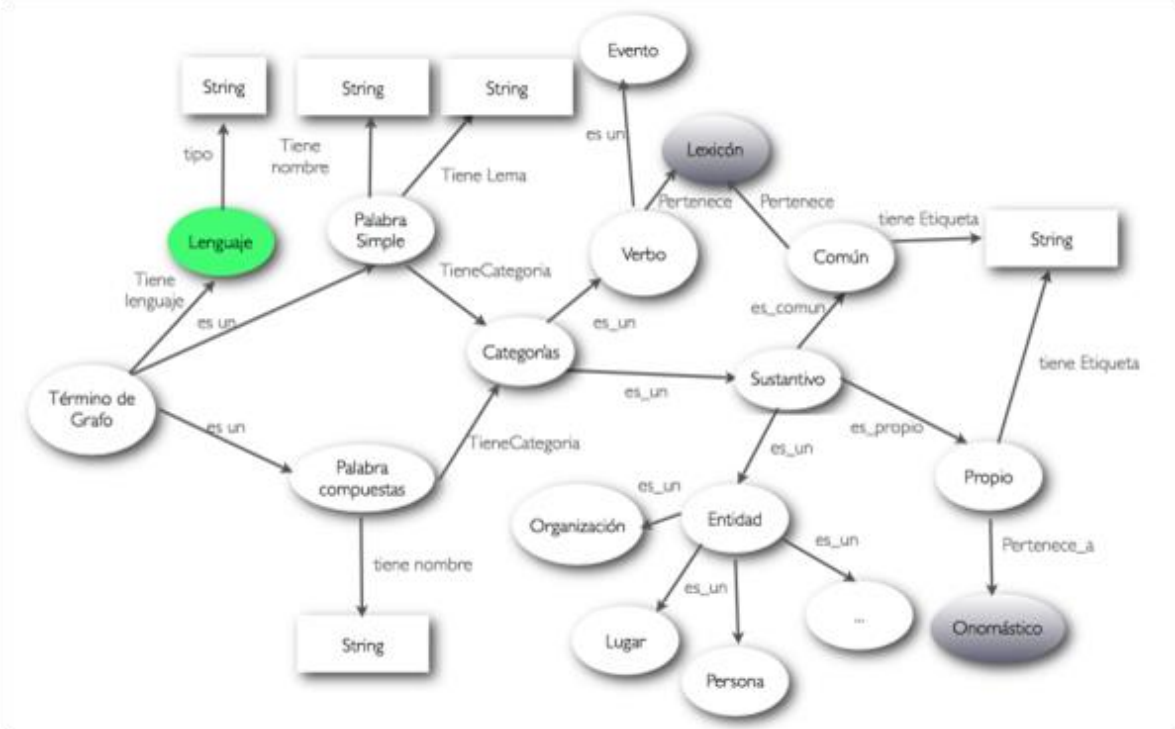


Figura 4.5 Esquema conceptual del Sistema de Tratamiento del texto

4.4.2. Obtención de enlaces/documentos relevantes

En este proceso se realiza una clasificación, la cual consiste en obtener un conjunto de documentos recuperados como documentos relevantes para MODS (ver Figura. 4.6). Este proceso se basa en el Sistema para la extracción de Documentos relevantes desde documentos recuperados en la Web, presentado en [44]. Su objetivo es tomar los documentos recuperados por cualquier buscador de la Web, y en base a la consulta realizada por el usuario o por un sistema de interpretación de la consulta, que en este caso es el MODS, y la definición de criterios para la selección de documentos relevantes, este sistema usa un modelo vectorial para identificar los documentos relevantes. A continuación se describe el proceso general (ver Figura. 4.6):

- ✓ El buscador da como salida un conjunto de documentos, llamados documentos candidatos, que supuestamente satisfacen la necesidad de información del MODS.
- ✓ Este listado de enlaces candidatos es filtrado usando los criterios de relevancia especificados, para definir la entrada al proceso de obtención de los documentos relevantes. Esos criterios de relevancia son previamente definidos por el MODS. Entre los criterios de relevancia se tienen, por ejemplo, el tipo de documento (html, pdf, owl, rdf, etc), su lenguaje (español, inglés, etc.), si los enlaces candidatos están inactivoso no[62], entre otros. Si los enlaces no cumplen con los criterios de relevancia definidos, dejan de ser enlaces candidatos.
- ✓ Finalmente, se realiza el proceso de identificación de los documentos relevantes, en el que se explora a cada documento candidato usando la técnica de recuperación de información basada en el modelo vectorial (ver capítulo 2). Esta técnica nos permite determinar la similitud de los documentos con la consulta, dando como resultado los documentos relevantes. En específico, para obtener los documentos relevantes se realiza los siguientes pasos (ver Figura. 4.6):
    - Análisis de la frecuencia, el cual consiste en contabilizar el número de ocurrencias de los términos que se encuentra en la consulta y en los documentos recuperados.
    - Luego se obtienen los pesos TF-IDF, que consiste en el cálculo de la importancia de un término para discriminar un documento, o una colección de ellos. Recordemos que para hallar esos pesos se debe previamente definir la frecuencia inversa (IDF). Las fórmulas matemáticas para obtener esas variables fueron definidas en el capítulo 2, ecuaciones 2.1 y 2.2, basadas en [32].
    - La frecuencia, como los pesos, son usados para el cálculo de la similitud entre la consulta y los documentos recuperados. La fórmula de similitud a usar fue presentada en el capítulo 2, y está basada en [7] (ver ecuación 2.3).

[62] Se consideran enlaces inactivos cuando se presentan los siguientes casos: cuando no logra localizar el documento, cuando el servidor no responde, cuando el acceso de la página está prohibido o se requiere una clave de acceso, cuando la página ha sido eliminada o trasladada a otro servidor, etc.

- Los documentos relevantes son ordenados de mayor a menor de acuerdo a los resultados de similitud, y se selecciona como los documentos relevantes, todos los documentos cuyo valor de similitud son mayores o iguales al promedio de la similitud de los documentos.

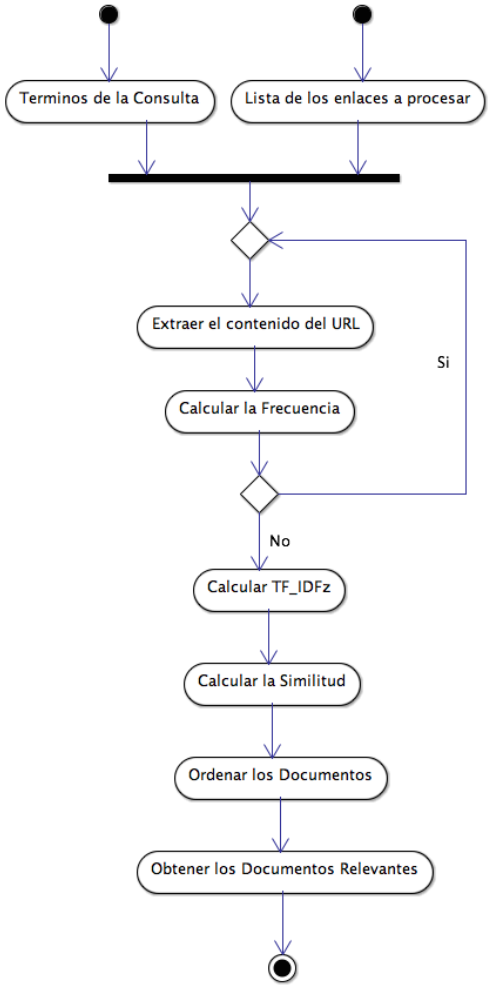


Figura. 4.6. Proceso de obtención de documentos (enlaces) relevantes.

Es de hacer notar que el modelo vectorial tiene como limitante la sensibilidad semántica, es decir que documentos con contextos similares pero con diferente vocabulario no serán asociados, dando falsos negativos. Para resolver esta limitante, en el caso concreto del MODS la consulta pasa por un proceso de interpretación, el cual le añade a la consulta aquellos términos relacionados que pueden utilizarse para expresar la misma idea o concepto (se extiende la consulta con contenido semántico relacionado a la misma).

4.4.3. Extracción de Conocimiento

La extracción de conocimiento en los documentos no estructurados consiste en extraer las palabras de alto contenido semántico, tales como las entidades (las

unidades básicas de los documentos, como por ejemplo los lugares, las personas, las organizaciones, etc.), los nombres propios (las diferentes formas que se usan en los documentos para referirse a una entidad), y las relaciones que se establecen entre las entidades, no tomando en cuenta las palabras de bajo contenido semántico, tales como las categorías gramaticales artículos, preposiciones, adjetivos y adverbios. Este proceso se realiza sobre los documentos relevantes previamente obtenidos.

En la Figura. 4.7 se presenta el proceso general de extracción de entidades y relaciones en textos no estructurados. En dicho proceso aparece una estructura de datos llamada Grafo de Aprendizaje (GA), que tiene por finalidad estructurar los términos extraídos del texto en un grafo (ese grafo es una *ontología semántica* con la información recuperada desde los textos, más adelante retomamos su análisis, y tiene como estructura de base la definida en la Figura 4.5). El proceso de extracción de conocimiento va creando instancias en esa ontología con el conocimiento descubierto.

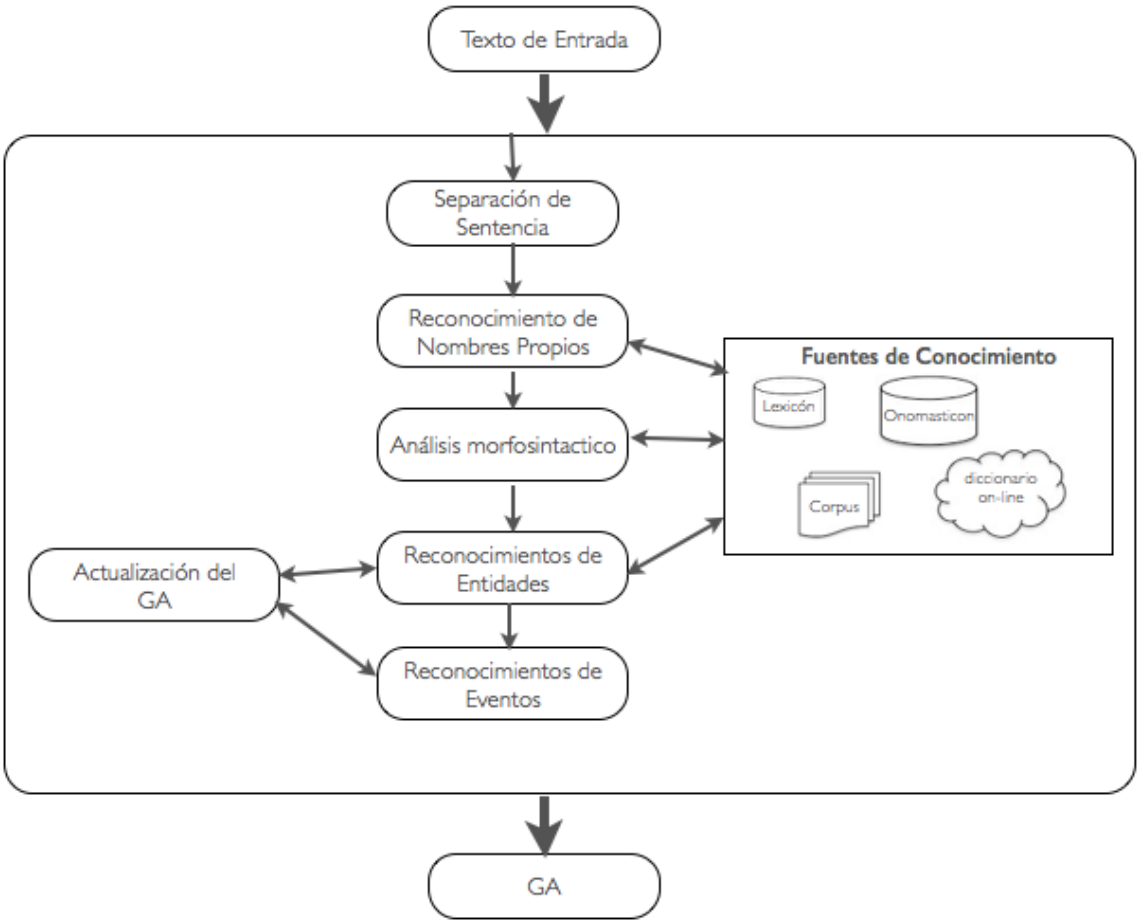


Figura 4.7. Proceso de extracción de conocimiento

El proceso se divide en 3 pasos, y uno de ellos en 4 sub-procesos, los cuales se describen a continuación [44]:

1. Separación del texto en Sentencias: este proceso divide el texto en sentencias.
2. Por cada Sentencia se realiza lo siguiente (análisis de las sentencias):
    a. Reconocimientos de Nombres Propios: consiste en la detección y clasificación de los términos del texto, como nombres de personas, organizaciones, lugares, expresiones numéricas, de tiempo, de lugar etc.; en este proceso se utiliza como recursos lingüísticos el Onomasticon y Corpus. El reconocimiento de nombres propios puede ser de palabras simples (como por ejemplo, Juan) o palabras compuestas (como por ejemplo, Universidad de Los Andes). Por ejemplo, la sentencia: “La Universidad de los Andes es una Universidad pública y autónoma”, identifica que Universidad de Los Andes es el nombre propio de Universidad.
    b. Análisis Morfosintáctico: Este análisis consiste en etiquetar las palabras de la sentencia con la información morfológica y sintáctica de ellas; es decir, cada una de las palabras de la sentencia se etiqueta según su categoría gramatical (verbos, sustantivos, adjetivos, verbos, etc.). Este proceso utiliza los recursos lingüísticos lexicón, onomasticon y Corpus.
    c. Reconocimiento de Entidades: en este proceso se extraen las palabras (tokens) etiquetadas como nombres propios y nombres comunes en la sentencia, en el análisis morfosintáctico. Ellas serán las entidades candidatas que serán usadas en el GA. En el reconocimiento de entidades simples se obtiene el lema de la entidad, el cual tiene como objeto normalizar las entidades (por ejemplo, las entidades simples Universitario y Universitarios tienen el lema Universitari, este lema representa la misma entidad)
    d. Reconocimientos de Relaciones: en este proceso se extraen las palabras (tokens) tipo verbos de la sentencia, etiquetadas en el análisis morfosintáctico. Ellas son las relaciones candidatas que serán usadas por el GA.
3. Actualización de GA: En este proceso se actualizan todas las entidades y relaciones candidatas encontradas en el texto procesado, con sus respectivas relaciones y tipos de datos, en el GA. Esta fase tiene características específicas que son indicadas en la siguiente sección.

4.4.4. Actualización del Grafo de Aprendizaje (GA)

GA permite describir ontológicamente (semánticamente) todo el conocimiento extraído desde el texto. De esta manera, GA es una ontología semántica que permite describir el conocimiento contenido en un texto (colección de textos). Sus instancias serán ese conocimiento extraído desde el (los) texto(s). Para ello, como dijimos anteriormente, GA se basado en el esquema conceptual mostrado en la Figura 4.5, y se define de las siguiente manera: un grafo GA es una tupla (T, E) donde:

- ✓ T es el conjunto de todos los conceptos involucrados en la caracterización de los términos encontrados en el texto
- ✓ R es el conjunto de sus relaciones

Para hacer ese proceso de caracterización de los términos en un texto, T contiene las siguientes clases (ver Figura. 4.8):

- ✓ Término Grafo: Representa todos los sustantivos y verbos que existen en el texto.
- ✓ Palabra Simples: es una subclase de Término Grafo, que representa todas las palabras atómicas.
- ✓ Palabra Compuesta: es una subclase de Término Grafo, que representa las palabras que están compuestas por varias palabras atómicas.
- ✓ Categoría: Representa la categoría gramatical del término Grafo, y está compuesta por Sustantivo y Verbo

Otras clases definidas en GA son los recursos lingüísticos en que se apoya este proceso (lexicón y onomasticon) y los tipos de sustantivos.

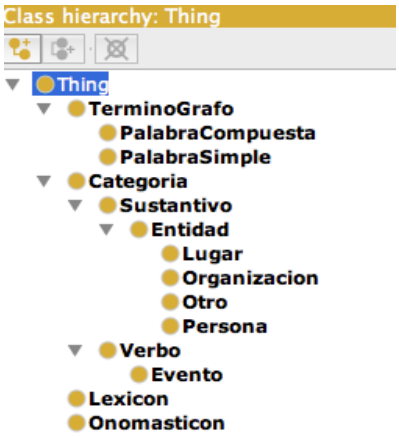


Figura. 4.8. Clases definidas en GA

Por otro lado, las relaciones entre esos conceptos son (ver Figura. 4.9):

- ✓ TieneGenero: indica si es femenino, masculino o neutro.
- ✓ TieneNumero: indica si el término es uno o más de uno. En específico, en español hay dos números: singular y plural.
- ✓ TieneSemantica: representa la clasificación semántica del término del grafo (si es organización, persona, lugar, otro)
- ✓ TieneTipo: representa el tipo del término (si es nombre común, nombre propio)
- ✓ esPalabraCompuesta: representa si el término está compuesto por varias palabras.
- ✓ esPalabraSimple: representa si el término es una sola palabra.

- ✓ esUn: representa si el término es un verbo
- ✓ tieneModo: En el caso de que el término sea un verbo, el modo se refiere a la actitud del hablante ante lo que dice. Hay tres modos verbales: Indicativo (por ejemplo, yo he llegado a la ciudad). Subjetivo (por ejemplo, quizás llegue a la ciudad) e Imperativo (por ejemplo, venid aquí).

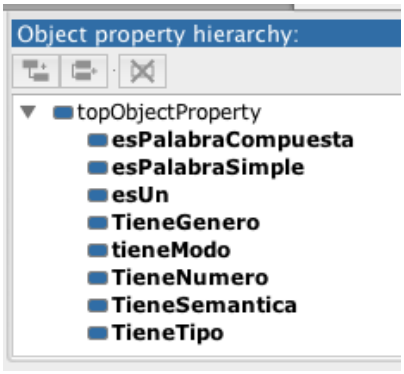


Figura 4.9. Relaciones definidas en GA

Finalmente, la clase término grafo (un término en GA) tiene un conjunto de propiedades (ver Figura. 4.10):

- ✓ tieneEtiquetaEAGLES: es una etiqueta que representa la información morfológica del término en el GA. Sirve para la anotación morfosintáctica de lexicones y corpus.
- ✓ tieneFrecuencia: representa el número de veces que se repite el término en el texto que se está procesando. Con esta frecuencia se determina si el término es relevante o no.
- ✓ tieneNombre: representa el nombre del término en el GA.

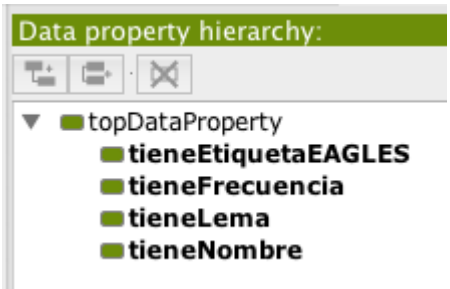


Figura. 4.10 Los tipos de datos definidos por defecto en el GA

De esta manera, cada instancia de la clase Término Grafo se describe de acuerdo a sus características morfológicas. Por ejemplo, el término Universidad se describe de la siguiente manera: Universidad es una palabra simple (relación esPalabraSimple), Universidad es femenino (relación TieneGenero), Universidad es singular (relación TieneNumero), Universidad es un nombre común (relación TieneTipo) y Universidad es una organización (relación TieneTipo).

En la Figura. 4.11 se muestra un ejemplo completo de cómo quedaría la descripción de la clase término grafo, en este caso para el término *Académico*, el cual se describe de la siguiente manera: es una palabra simple, de tipo nombre propio, singular, masculino, tiene el lema académ, se repite en los documentos dos veces, el nombre es Académico, y posee la etiqueta eagles np00000.

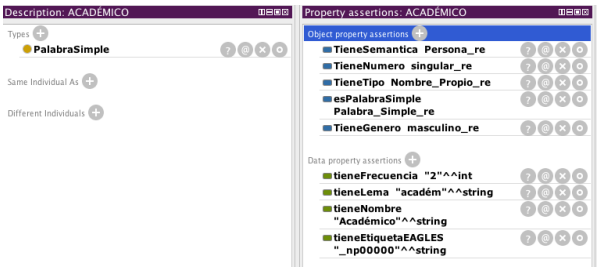


Figura. 4.11 Ejemplo de descripción de la clase Término Grafo instanciada en el término "Académico"

Retomemos de nuevo la descripción del proceso de aprendizaje semántico. Recordemos que en la fase de extracción de conocimiento se genera una colección de palabras (tokens) en un dominio específico, en los cuales se han caracterizado:

- Entidades Candidatas, se define como todos las sustantivos que se encuentran en los texto procesados.
- Relaciones Candidatas, se define como todos los verbos que se encuentran en los textos procesados.

En este paso se seleccionan las entidades y relaciones relevantes, que deberán ser incorporadas al GA (lo actualizan/instancian), las cuales se definen como:

- Entidades Relevantes, son todos los sustantivos que cumplen ciertos criterios de relevancia.
- Relaciones Relevantes, son todos los verbos que cumplen ciertos criterios de relevancias.

En este paso, al igual que en el paso de identificación de documentos relevantes, se requiere la definición de unos criterios para priorizar la información que se está obteniendo. Para este caso en específico, se proponen los siguientes criterios para clasificar a los términos como relevantes o no, inspirados en el trabajo [32]:

- ✓ tf (Frecuencia término): es la frecuencia de un término j. Es una medida de relevancia de un término en GA. Si el término aparece muchas veces en el GA el peso del término es alto.
- ✓ fi (frecuencia inversa): o inverso de la frecuencia de documentos, pretende reflejar la importancia de los términos en los textos procesados, primando el poder discriminatorio de los mismos. Así, dará mayor importancia a un término cuanto menor sea el número de texto procesados de la colección en los que aparezca dicho término. Por el contrario, si un término aparece en todos los

textos procesados, su precisión y poder discriminatorio será menor, de manera que se le otorgará una importancia mínima en esos textos procesados en concreto. Fi se calcula como:

$$fi_j = \log\left(\frac{Número \text{ de textos procesados}}{Número\ de\ texto\ donde\ aparece\ termino\ j}\right) \qquad (4.1)$$

- ✓ wj (peso del término j) indica que entre mayor es él para un término j, es porque ese término es más relevante, ya que es muy frecuente y discriminante.

$$w_j = \begin{cases} tf_j * fi_j & si\ textos\ procesados > 1 \\ fi_j & de\ lo\ contrario \end{cases} \qquad (4.2)$$

En concreto, el criterio utilizado para determinar si una Entidad o Relación es relevante es: será relevante si su peso es mayor o igual que el promedio de los pesos.

4.4.5. Actualización de Ontologías

Una vez obtenido GA, que es una ontología semántica, se procede a extraer la información necesaria para la fase de actualización. Esta fase consiste en actualizar los componentes del MODS, en este caso se procede actualizar a las ontologías lexicón, onomasticon e interpretativa. A continuación se describen los tres procesos de actualización.

Actualización del lexicón

Para el proceso de actualización de la ontología del lexicón se realiza lo siguiente (ver Figura 4.12):

1. Extraer los Sustantivos y Verbos de la Ontología Semántica: este proceso consiste en obtener todos los sustantivos y verbos contenidos en la ontología semántica (GA), por medio de la siguiente regla:

TieneTipo(?x, Nombre_Comun_re), Peso(?x, ?fre), greaterThan(?fre, ?prompeso) -> Lexicon(?x).

esUn(?x, Verbo_re), Peso(?x, ?fre), greaterThan(?fre, ?prompeso) -> Lexicon(?x)

2. Hacer por cada sustantivo y verbo lo siguiente

a. Buscar el término en el lexicón y si lo encuentra:
   i. Actualizar las relaciones
   ii. Actualizar las propiedades

   En caso contrario
   Insertar el término con sus relaciones y propiedades

Actualización del onomasticon

Desde la ontología semántica se pueden obtener los nombres propios para actualizar el onomasticon. Para la actualización del onomasticon se realiza el mismo proceso de la Figura 4.12, la única diferencia es que se extraen los

nombres propios de la ontología semántica, utilizando la siguiente consulta SPARQL:

TieneTipo(?x, Nombre_Propio_re), Peso(?x, ?fre), greaterThan(?fre, prompeso) -> Onomasticon(?x)

Actualización de la Ontología Interpretativa

Para realizar la actualización de la Ontología Interpretativa, desde la Ontología Semántica se obtienen:

1. Las entidades relevantes
2. Las relaciones relevantes
3. Los textos donde se encuentra las entidades y las relaciones

Con esta información se infieren los conceptos, y las relaciones taxonómicas y no taxonómicas. Por ejemplo, los patrones que se buscan en los textos para determinar las relaciones taxonómicas son del tipo: “X es un Y”, “X es un tipo de Y”, “ X de cierta manera es Y”, “X en cierto sentido es Y”. En el caso de las relaciones no taxonómicas, para el caso de sinónimos, los patrones usados son del tipo: “X es denominado Y”, “X es equivalente a Y”, “X es igual que Y”, “X es idéntico a Y”, “X es sinónimo de Y”. Otros ejemplos de patrones para otras relaciones no taxonómicas usados en este trabajo, son dados en el apéndice C.

En particular, esta actualización básicamente es un proceso de mezcla ontológica [26], que consiste en generar una única ontología a partir de las ontologías originales. En nuestro caso, el proceso de mezcla consiste en generar una única ontología interpretativa a partir de la ontología semántica y la ontología interpretativa actual. Es decir, la ontología interpretativa actual es actualizada con los conceptos, relaciones y tipos de datos presentes en GA.

Para realizar ese proceso en nuestro trabajo, nosotros usamos el siguiente procedimiento (ver Figura. 4.12): vamos analizando el contenido en GA y verificando si está contenido en la ontología interpretativa.

Caso 1. Actualizar Relaciones Taxonómicas.

1. Se realiza la búsqueda de las entidades X e Y contenidas en GA, en la ontología interpretativa:
    a. en el caso que encontró las dos entidades se verifica si X es subclase de Y. En el caso que no exista la subclase se actualiza SubClassOf(X,Y) en la ontología interpretativa.
    b. En el caso que encontró X y no se encontró Y. Se agrega la entidad Y en la ontología interpretativa, y se agrega la relación SubClassOf(X,Y), el mismo procedimiento para el caso que no encontró X.
    c. En el caso de que no encontró X e Y. Se agrega la entidad X e Y en la ontología interpretativa, y se agrega la relación SubClassOf(X,Y)

Caso 2. Actualizar Relaciones no-Taxonómicas.

1. Se realiza la búsqueda de la relación Z y las entidades X e Y contenidas en GA, en la ontología interpretativa:
    a. En caso que encontró la relación Z, se verifica si existe las dos entidades:
        i. Si existe las dos entidades, se verifica si están relacionadas con la relación Z, si no están relacionadas se relacionan las dos entidades.
        ii. Si no existe la entidad X y existe la entidad Y, se agrega la entidad X y se agrega la relación_Z(X,Y), de igual manera en el caso que no exista Y
        iii. En el caso que no existe la entidad X e Y, se agrega las entidades X e Y, se agrega la relación_Z(X,Y).
    b. En caso que no encontró la relación Z, se verifica si existe las dos entidades:
        i. Si existe las dos entidades, se agrega la relación_Z(X,Y)
        ii. Si no existe la entidad X y existe la entidad Y, se agrega la entidad X y se agrega la relación_Z(X,Y), de igual manera en el caso que no exista Y
        iii. En el caso que no existe la entidad X e Y, se agrega las entidades X e Y, se agrega la relación_Z(X,Y).
        iv.

Caso 3. Actualización de Propiedades.

1. En el caso que las entidades X e Y contenidas en GA, existan en la ontología Interpretativa
    a. Se actualiza las propiedades de cada una de las entidades.

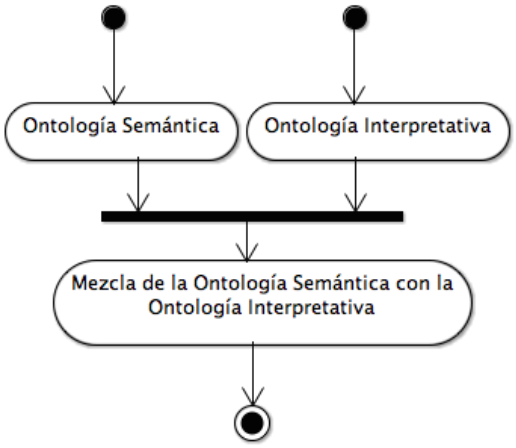


Figura 4.12 Proceso de actualización de la ontología interpretativa

4.5. Conclusión

En este capítulo se presentó la arquitectura de aprendizaje ontológico general del MODS, y en detalle cada uno de sus componentes. La arquitectura es capaz

de adquirir nuevo conocimiento en función de la información suministrada por el proceso de análisis de la consulta en lenguaje natural, y/o de la información recuperada desde la Web por dicha consulta. Esta arquitectura, a partir del nuevo conocimiento adquirido, permite potenciar las estructuras del MODS.

La unidad de aprendizaje morfosintáctico permite diferenciar entre un verbo regular e irregular, y conjugarlo en todas las formas personales e impersonales en varios tiempos (aprendizaje duro). Además, cuenta con otros componentes (el aprendizaje simiente y el aprendizaje ágil), que le permiten aprender otras categorías lingüísticas: sustantivos, adjetivos y adverbios.

La unidad de aprendizaje semántico se caracteriza por integrar diferentes métodos y técnicas de descubrimiento y comparación de conocimiento léxico, morfológico, sintáctico, semántico y pragmático.

Además se diseñaron otros módulos, para la unidad de aprendizaje semántico un sistema de extracción de documentos relevantes sobre los resultados ya recuperados por cualquier buscador en Internet, tomando en cuenta un conjunto de criterios que define el usuario y su consulta realizada. Este sistema es usado por el MODS para la extracción de conocimiento desde los documentos recuperados. Este sistema tiene como característica que el proceso de filtrado se realiza analizando el texto completamente, y no solamente tomando fragmentos de textos (muestras), desde lo que retorna el motor de búsqueda; además, con la expansión de la consulta realizada por el MODS nos garantiza resolver el problema de falsos negativos. El cálculo de la similitud nos garantiza la obtención de documentos relevantes con un 100% de precisión, algo tampoco garantizado en los trabajos previos analizados. El otro modulo diseñado es el sistema de extracción de conocimiento desde textos no estructurados, el cual es organizado semántica en un Grafo de Aprendizaje (GA). GA es una ontología semántica que puede ser representada en RDF, OWL, etc., cuyas instancias representan todo el conocimiento extraído de una colección de documentos. Para la construcción del GA hemos usado una métrica que permite clasificar los términos contenidos en los documentos procesados como relevantes o no. Dicha métrica combina la frecuencia y la capacidad de discriminación de los términos en los documentos procesados. Esto le confiere al GA una característica fundamental, ser una ontología semántica de documentos procesados, conformada solamente por información (individuos) relevantes. Trabajos previos no consideran estos aspectos, es decir el desarrollo de ontologías semánticas que contienen todo el conocimiento relevante de un conjunto de documentos bajo estudio. Este sistema es usado como insumo por la unidad del aprendizaje semántico del MODS. Este sistema permite actualizar los siguientes componentes del MODS: la ontología interpretativa, el onomasticon y el lexicón. GA da el insumo de las entidades y relaciones a incorporar en la ontología interpretativa, define los nuevos conceptos (verbos y sustantivos) a incorporar en el lexicón, y los nuevos nombres propios a incorporar en el onomasticon.

Es de hacer notar que este proceso se hace después que el usuario obtiene la respuesta, y es usado solo para las actualizaciones de los componentes del MODS. El rendimiento del sistema depende de varios factores, de la conexión de

internet, del total de documentos a procesar, de las características del equipo. Además el proceso no es totalmente automático, sino semiautomático, porque se requiere del experto para validar la información que se está agregando en cada componente (aunque esta fase de validación se puede obviar si se tiene confianza en la fuente de información de la Web).

Capítulo 5: Implementación y experimentación con el MODS

En este capítulo se presenta la arquitectura de implementación de los componentes del MODS (cada componente usa específicas librerías: OpenNLP[63], Apache Jena[64], OWLAPI[65], etc.). Después se presentan los casos de uso del sistema cuando todos los términos de la consulta son conocidos. Seguidamente se presenta cuando algún término de la consulta no es un término conocido, por lo cual se invoca el componente de aprendizaje morfosintáctico del MODS. A continuación, se presenta cuando la consulta es realizada con éxito, y se llama el componente de aprendizaje semántico para que actualice las ontologías del MODS. Después se presentan otros casos de estudio de minería semántica, para finalmente realizar un análisis cualitativo de todos los resultados obtenidos en este trabajo.

5.1. Arquitectura del MODS.

La arquitectura de implementación del MODS se presenta en la Figura. 5.1.

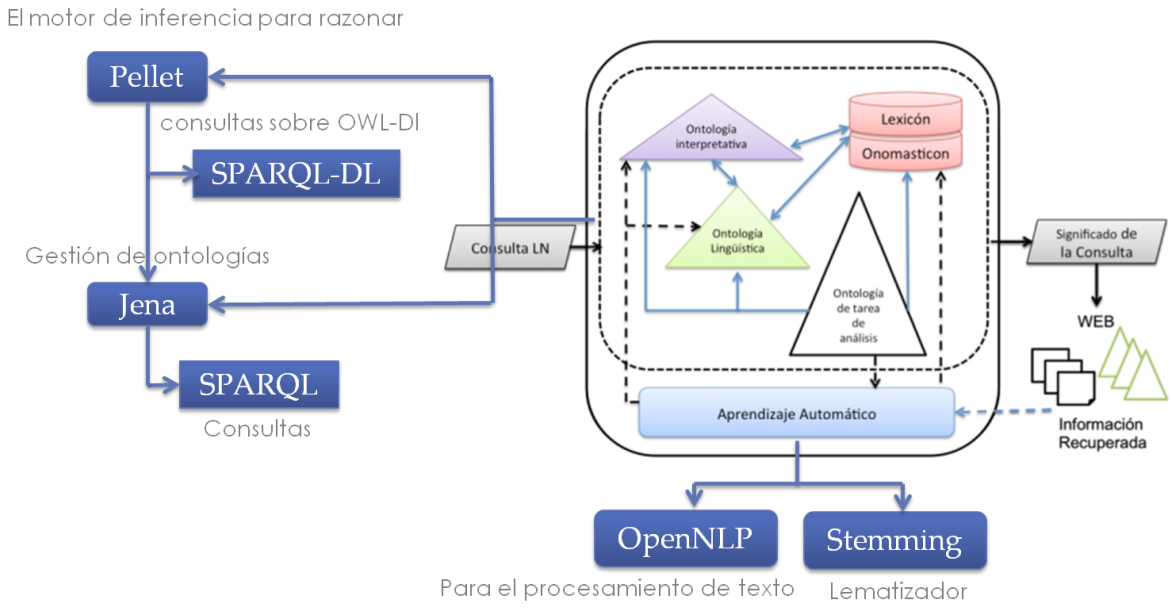


Figura. 5.1. Arquitectura de Implementación del MODS

[63] OpenNLP es un proyecto de la Apache Software Foundation, de uso libre y open source, que contiene una suite de herramientas para el procesamiento de lenguaje natural. Su dirección electrónica es http://opennlp.apache.org/
[64] https://jena.apache.org/
[65] http://owlapi.sourceforge.net/

En general, se usa al editor de ontologías protégé[66], para invocar a muchas de esas librerías. En particular, en la Tabla 5.1. se muestra que componentes del MODS usan que API:

Tabla 5.1 API utilizados para la implementación del MODS

API	Descripción
Jena	Este API se usa para la manipulación de todas las ontologías del MODS, por ejemplo, para crear, leer o actualizar las ontologías y el grafo de aprendizaje (GA). También se utiliza su motor de consultas, compatible con SPARQL
OpenNLP	Este API se usa para el procesamiento de textos, en el componente de aprendizaje del MODS.
Pellet	Este API se usa como motor de inferencia, para razonar sobre las ontologías del MODS.
OWL	Este API se usa para la manipulación de las ontologías basadas en OWL.
Stemming	Es el lematizador utilizado por el componente de aprendizaje del MODS

5.2. Consulta con todos los términos conocidos

En este primer caso se realiza una consulta al MODS donde todos los términos son conocidos. Usaremos el ejemplo: “Universidad de Los Andes”, en este caso el MODS sigue el proceso que se muestra en la Figura. 5.2 (para lo cual usa la ontología de tareas).

[66] http://protege.stanford.edu/

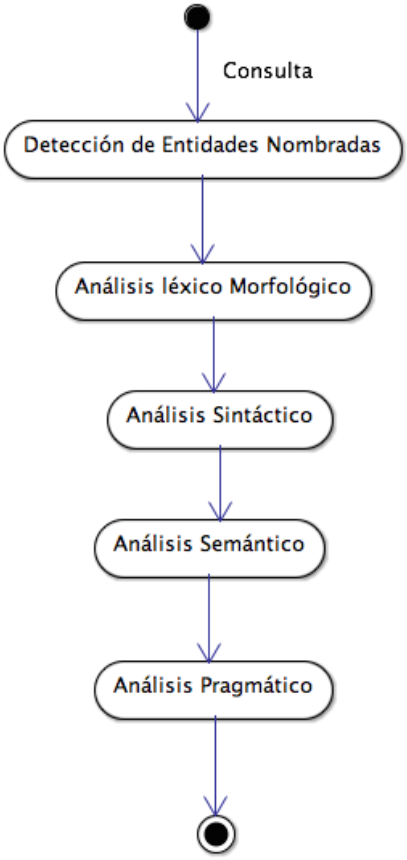
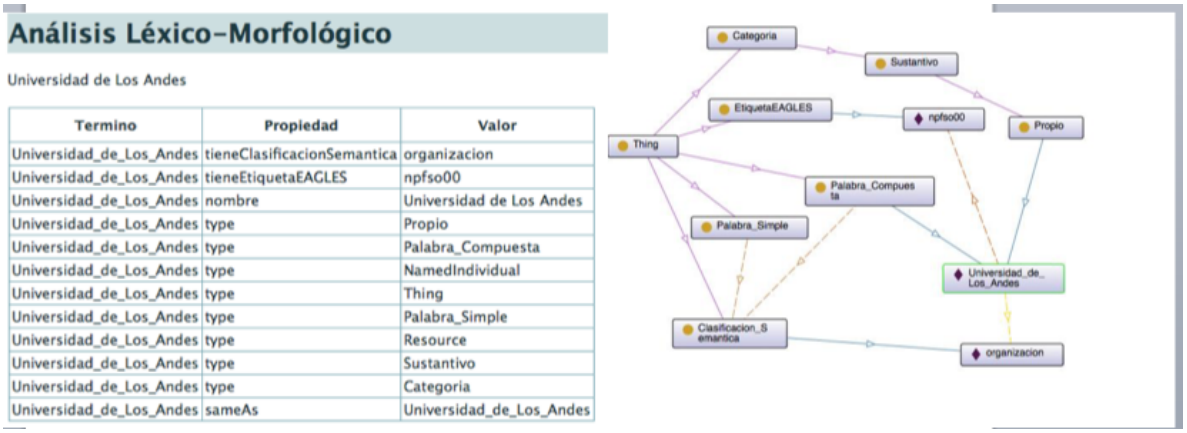


Figura 5.2. Proceso de la consulta sobre MODS cuando todos los términos son conocidos

El proceso comienza con la detección de las entidades nombradas en la consulta. En el caso del ejemplo se encontró que Universidad de Los Andes es una entidad nombrada. Siguiendo con el proceso, se realiza el análisis léxico-morfológico, en la Figura. 5.3 se muestra el resultado de este proceso. Allí básicamente se determina que la Universidad de Los Andes es un nombre propio, de categoría sustantivo, es una palabra compuesta, y tiene etiqueta EAGLES npfso00; esta información es obtenida desde el Onomasticon.



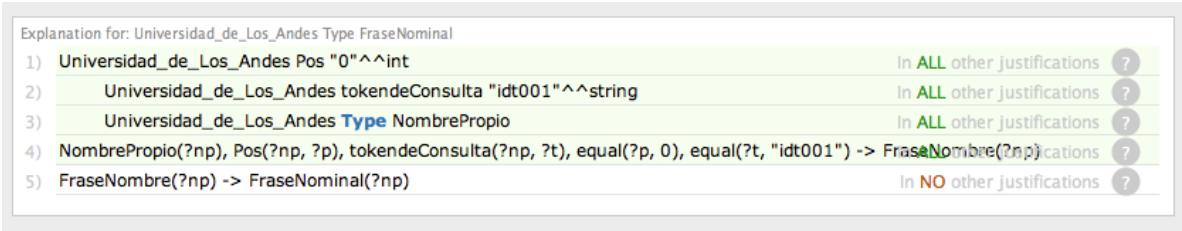
Análisis Léxico-Morfológico

Universidad de Los Andes

Termino	Propiedad	Valor
Universidad_de_Los_Andes	tieneClasificacionSemantica	organizacion
Universidad_de_Los_Andes	tieneEtiquetaEAGLES	npfso00
Universidad_de_Los_Andes	nombre	Universidad de Los Andes
Universidad_de_Los_Andes	type	Propio
Universidad_de_Los_Andes	type	Palabra_Compuesta
Universidad_de_Los_Andes	type	NamedIndividual
Universidad_de_Los_Andes	type	Thing
Universidad_de_Los_Andes	type	Palabra_Simple
Universidad_de_Los_Andes	type	Resource
Universidad_de_Los_Andes	type	Sustantivo
Universidad_de_Los_Andes	type	Categoria
Universidad_de_Los_Andes	sameAs	Universidad_de_Los_Andes

Figura 5.3. Tarea de Análisis léxico-Morfológico

El sistema forma la siguiente regla: *Universidad_de_Los_Andes*_npfso00.El proceso sigue con el análisis sintáctico, el cual consiste en generar el grafo sintáctico de la consulta. En la Figura. 5.4. se muestra el resultado de este proceso en protégé.



Explanation for: Universidad_de_Los_Andes Type FraseNominal

1) Universidad_de_Los_Andes Pos "0"^^int — In ALL other justifications
2) Universidad_de_Los_Andes tokendeConsulta "idt001"^^string — In ALL other justifications
3) Universidad_de_Los_Andes **Type** NombrePropio — In ALL other justifications
4) NombrePropio(?np), Pos(?np, ?p), tokendeConsulta(?np, ?t), equal(?p, 0), equal(?t, "idt001") -> FraseNombre(?np) — In ALL other justifications
5) FraseNombre(?np) -> FraseNominal(?np) — In NO other justifications

Figura 5.4. Resultado de Análisis sintáctico usando la ontología lingüística

En este caso se ejecutan las reglas respectivas de la ontología lingüística. Las mismas consisten, para el caso de estudio, en (ver Figura. 5.4):

NombrePropio(Universidad_de_Los_Andes)[67] -> FraseNombre[68]
FraseNombre(Universidad_de_Los_Andes) -> FraseNominal[69]

Luego se sigue con el proceso de análisis semántico, la cual considera la estructura semántica de la regla sintáctica, tal como se muestra en la Figura. 5.5. Básicamente, para el caso de estudio, se muestra que el nombre propio “Universidad_de_Los_Andes” es un sujeto, y su estructura semántica es que el sujeto es agente, aplicando las siguientes reglas:

NombrePropio(Universidad_de_Los_Andes) -> Sujeto
Agente->Subjeto(Universidad_de_Los_Andes) -> Agente

[67] NombrePropio -> (Np)
[68] FraseNombre es equivalente a Sintagma nombre (SNo)
[69] FraseNominal es equivalente a Sintagma nominal (SN)

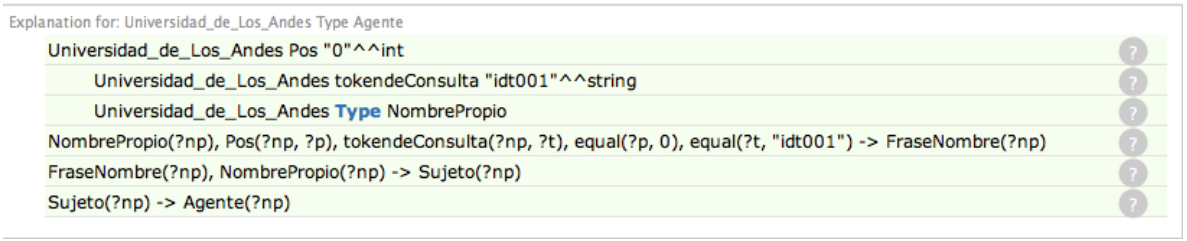


Figura. 5.5. Resultado de Análisis semántico usando la ontología lingüística

Como conclusión de ese proceso obtenemos que la Universidad de Los Andes es un agente. Luego, continua el proceso de análisis pragmático, el cual se apoya en la ontología interpretativa para indicar, además, que la Universidad de Los Andes es una Universidad que está ubicada en Venezuela, en la ciudad de Mérida, generando la siguiente consulta booleana (es una consulta extendida de la original):

(“Universidad de los Andes” and Mérida and Venezuela) or (ULA Mérida and Venezuela) or (“Universidad de los Andes” and “Núcleo Mérida” and Mérida and Venezuela) or (ULA and “Núcleo Mérida” and Mérida and Venezuela)

Este proceso es el que sigue el MODS cuando la consulta tiene todos los términos conocidos. En el caso que no se conoce un término de la consulta, la ontología de tarea invoca al componente de aprendizaje, el cual se explica en la siguiente sección.

5.3. Componente de Aprendizaje morfosintáctico: se invoca cuando no se conoce un término de la consulta

En esta sesión se analiza el caso de que un término de la consulta sea desconocido. El término desconocido puede ser un sustantivo, verbo, adjetivo o un adverbio. Para todos los ejemplos, se parte de la siguiente estructura:

Lex_mor(Componente léxico, categoría, tipo, genero, numero, modo, tiempo, aspecto, voz, persona)

Esta estructura es la interfaz entre el resto de componentes del MODS y el componente de aprendizaje, tal como se define en la sección 4.3 del capítulo 4. A través de ella se invoca al componente de aprendizaje cuando hay palabras desconocidas en la consulta, instanciando esa estructura con el término desconocido. El componente de aprendizaje retorna esa estructura con la información encontrada, según la categoría lingüística del término desconocido. A continuación, se muestra un ejemplo para el aprendizaje de sustantivos, adjetivos, verbos y adverbio.

5.3.1. Aprendizaje de un sustantivo

Para el caso de la siguiente entrada:

lex_morf(cama, Desconocido, null, null, null, null, null, null,null, null)

El componente de aprendizaje invoca al proceso de aprendizaje morfosintáctico explicado en el capítulo 4. A groso modo, el proceso es el siguiente:

1. El aprendizaje simiente determina cual es la forma canónica de la palabra “cama”. La respuesta es: cama tieneUnlema cama.
2. El aprendizaje ágil consulta al diccionario WordReference.com. Su respuesta es que: cama tieneCategoria Sustantivo, cama tieneGenero femenino, cama tipoSustantivo común y cama tieneEtiquetaEAGLES ncfs000.
3. Se instancia al grafo de aprendizaje morfosintáctico con esa información (ver Figura. 5.6).
4. Se actualiza el lexicón.

Ese mismo proceso se realiza para el caso de adjetivos y adverbios desconocidos. Ejemplos de instanciación del grafo de aprendizaje, derivados del proceso de aprendizaje morfosintáctico, para el caso de aprender el sustantivo casa, el adjetivo amable, y el adverbio abajo, son mostrados en la Figura 5.6.

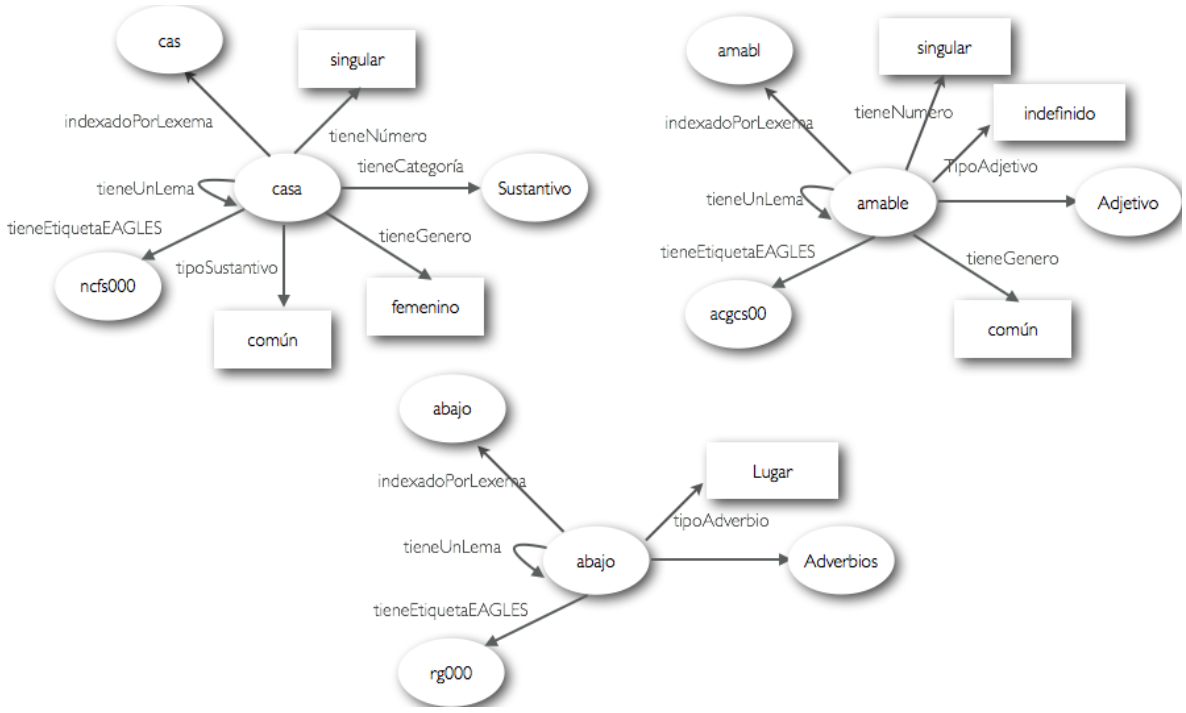


Figura 5.6. Grafo de aprendizaje para el caso de aprender el sustantivo casa, el adjetivo amable y el adverbio abajo.

5.3.2. Aprendizaje de un verbo

Si se supone ahora la siguiente entrada:

lex_morf(compro, Desconocido, null, null, null, null, null, null,null, null)

El proceso del aprendizaje morfosintáctico es el siguiente:

1. El aprendizaje simiente determina cual es la forma canónica, en el caso de un verbo es su infinitivo. La respuesta en este caso es: compro tieneUnlema comprar.
2. El aprendizaje ágil consulta al diccionario WordReference.com. Su respuesta es que: compro tieneCategoria Verbo, compro tieneModo Indicativo, compro tipoVerbo Transitivo, compro tieneTiempo Presente, compro tienePersona Primera y compro tieneEtiquetaEAGLES vmipls0.
3. Este resultado es incorporado al grafo de aprendizaje morfosintáctico (ver Figura. 5.7).
4. El aprendizaje duro, realiza la conjugación de acuerdo al tipo de verbo (regular o irregular), dando la siguiente salida para el caso del verbo comprar y en el término compraba (para el resto de conjugaciones: compraré, compré, compramos, etc., es similar): compraba tieneCategoria Verbo, compraba tieneModo Indicativo, compraba tipoVerbo Transitivo, compraba tieneTiempo Imperfecto, compro tienePersona Primera y compraba tieneEtiquetaEAGLES vmipls0, etc.
5. Se instancia al grafo de aprendizaje con esa información (ver Figura 5.7).
6. Se actualiza el lexicón

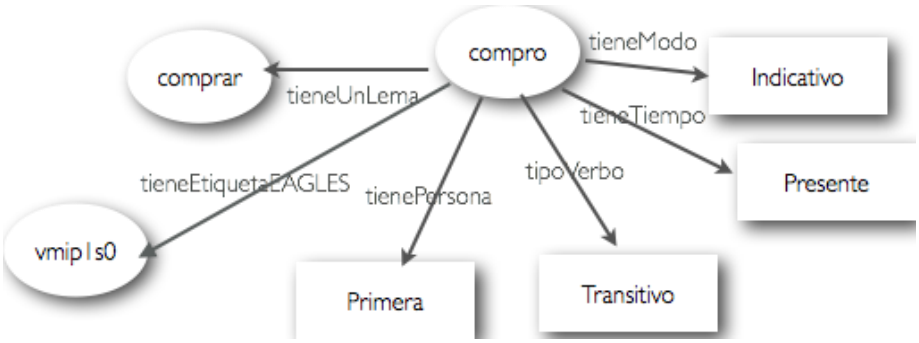


Figura 5.7. Grafo de aprendizaje de verbo comprar.

5.4. Invocación al componente de aprendizaje semántico

El componente de aprendizaje semántico se invoca cuando el proceso de interpretación de la consulta es exitoso, lo que resulta en un grupo de información recuperada desde la Web. A partir de esa información, el MODS invoca al componente de aprendizaje semántico, para actualizar la ontología interpretativa.

En este aprendizaje se van a extraer los documentos no estructurados (html) relevantes para la consulta, para luego extraer de ellos el conocimiento requerido, en nuestro caso entidades y relaciones.

Recordemos que el proceso el aprendizaje semántico es invocado por la ontología de tareas, después de pasar por el proceso de interpretación de la consulta del MODS. Ese proceso de interpretación de la consulta genera la siguiente consulta booleana (ver en la sección 5.2 cómo se obtiene):

("Universidad de los Andes" and Mérida and Venezuela) or (ULA Mérida and Venezuela) or ("Universidad de los Andes" and "Núcleo Mérida" and Mérida and Venezuela) or (ULA and "Núcleo Mérida" and Mérida and Venezuela)

Esa consulta extendida es realizada por el MODS en el buscador Google, y se obtiene como resultado un conjunto de enlaces a documentos (direcciones URL). Para explicar el proceso el aprendizaje semántico, se van a tomar los primeros 173 enlaces devueltos.

Siguiendo el procedimiento indicado en la sección 4.4.2. del capítulo 4, se seleccionan los enlaces candidatos bajo los siguientes criterios: documentos de tipo html, lenguaje del documento español, se eliminan todos los enlaces que comience con https o que terminen con extensiones pdf, jpg, etc. También, se eliminan los enlaces que direccionan a aplicaciones y a motores de búsquedas, o que no estén activos en el momento de la consulta.

Una vez obtenido la lista de los enlaces candidatos, se realiza el proceso de obtención de documentos relevantes basado en lo descrito en la sección 4.4.2. En particular, consiste en los siguientes pasos:

1. Se forma el vector de consulta, para lo cual se descompone la consulta en cuatro sub-consultas, que se muestran a continuación:

   q_1= "Universidad de Los Andes" and Mérida and Venezuela
   q_2= "ULA" and Mérida and Venezuela
   q_3= "Universidad de Los Andes" and "Núcleo Mérida" and Mérida and Venezuela
   q_4= "ULA" and "Núcleo Mérida" and Mérida and Venezuela

Esas son las posibles combinaciones de las palabras existentes en la consulta original del usuario, procesada/interpretada por el MODS. Luego, se obtiene el vector de consulta, que indica la frecuencia con la que aparece cada palabra en la consulta (ver la Tabla 5.2).

Tabla 5.2. Matriz de la consulta realizada por el usuario después que pasa por el proceso de interpretación del MODS

Consulta	Universidad de Los Andes	Mérida	Venezuela	ULA	Núcleo Mérida
q_1	1	1	1	0	0
q_2	0	1	1	1	0
q_3	1	1	1	0	1
q_4	0	1	1	1	1

2. Ahora se procesan los documentos a los que enlaza la lista de los enlaces candidatos, para obtener los documentos relevantes (ver la sección 4.4.2, donde se indican los detalles):

    a. Se procede a generar la matriz de frecuencia de los términos en los documentos, para ello se recogen las apariciones de cada término de la consulta en cada documento que se esta procesando.
    b. Se procede a calcular la frecuencia inversa de cada uno de los términos en cada documento (ver ecuación 2.2).
    c. Luego se calculan las similitudes existentes entre los distintos enlaces y la matriz de la consulta extendida (Tabla. 5.2), basada en la ecuación 2.3.

La Figura. 5.8. muestra el resultado de todo ese proceso sobre los documentos candidatos (para los primeros 34 enlaces), con el valor de similitud de cada documento con respecto a la consulta extendida.

Doc	Enlace	Similitud
73	http://erevistas.saber.ula.ve/index.php/visiongere	0,97
26	http://www.diarioeltiempo.com.ve/V3_Secciones/	0,93
37	http://www.luz.edu.ve/index.php?option=com_co	0,93
103	http://www.ula.ve/index.php?option=com_conter	0,93
33	http://www.infoandina.org/content/postgrado-en	0,92
83	http://www.saber.ula.ve/handle/123456789/3529	0,92
98	http://saber.ula.ve/dspace/handle/123456789/15	0,92
32	http://www.noticiasdiarias.informe25.com/2014/0	0,88
61	http://www.monografias.com/trabajos93/factores	0,87
50	http://aulaverde.masverdedigital.com/?p=2433	0,86
3	http://www.slideshare.net/MANUELLITOR	0,85
100	http://saber.ula.ve/handle/123456789/9510?mod	0,82
94	http://saber.ula.ve/handle/123456789/8614?mod	0,81
4	http://www.noticias24.com/venezuela/noticia/221	0,79
29	http://www.entornointeligente.com/articulo/2040	0,79
1	http://es.wikipedia.org/wiki/Universidad_de_Los_	0,77
68	http://estudianteselectrica.blogspot.com/	0,77
80	http://saber.ula.ve/handle/123456789/8417?mod	0,77
90	http://saber.ula.ve/handle/123456789/10022?mo	0,77
51	http://agenda.universia.edu.ve/ula/2005/03/18/p	0,76
79	http://saber.ula.ve/handle/123456789/10054?mo	0,76
91	http://saber.ula.ve/handle/123456789/8426?mod	0,75
87	http://saber.ula.ve/handle/123456789/8452?mod	0,74
99	http://saber.ula.ve/handle/123456789/10046	0,74
84	http://saber.ula.ve/handle/123456789/10004?mo	0,73
86	http://saber.ula.ve/handle/123456789/9082?mod	0,73
101	http://saber.ula.ve/handle/123456789/10007?mo	0,73
39	http://es.mashpedia.com/Universidad_de_Los_An	0,71
75	http://saber.ula.ve/handle/123456789/8221?mod	0,70
31	http://library.kiwix.org/wikipedia_es_venezuela_1	0,68
62	http://eurolobo.blogspot.com/2011_02_01_archiv	0,68
70	http://diccionario.sensagent.com/universidad+de+	0,68
82	http://saber.ula.ve/handle/123456789/8867?mod	0,67

Figura. 5.8. Lista de los enlaces relevantes

La tercera columna de la Figura. 5.8. expresa el valor de similitud entre los documentos y las consultas. En particular, muestra los enlaces más relevantes (documentos relevantes), ordenados de mayor a menor. Si se supone que las similitudes mayores al valor promedio de las similitudes son los

enlaces/documentos más relevantes, en este caso serán los mayores o iguales a 0,44.

Una vez obtenido los enlaces relevantes, se comienza el proceso de extracción de conocimiento desde ellos, el cual consiste en obtener las entidades y relaciones relevantes. Recordemos que para ello se sigue el procedimiento definido en la sección 4.4.3 (ver Figura. 4.7), el cual consiste en:

1. Por cada enlace se procesa el texto, para lo cual se separan en sentencias.
2. Luego se analizan las sentencias, para entre otras cosas, reconocer los nombres propios, las entidades y relaciones candidatas, presentes en cada una.

Finalmente se instancia el grafo de aprendizaje (GA) con las entidades y relaciones relevantes, extraídas desde las candidatas (la Figura. 5.9 muestra la instanciación del GA para este caso de estudio).

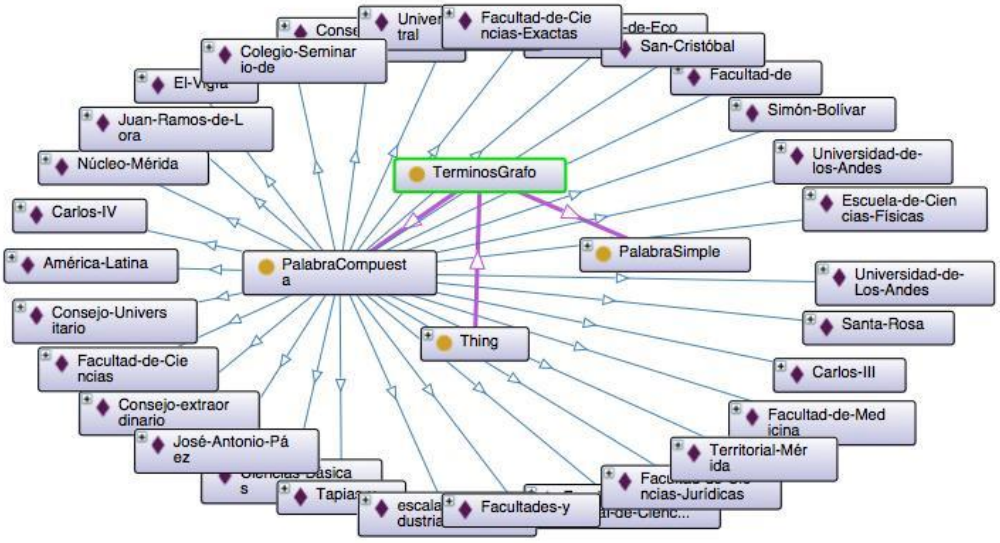


Figura. 5.9. Grafo de aprendizaje.

Recordemos que el criterio utilizado para determinar si una Entidad o Relación es relevante es (ver sección 4.4.4): “será relevante si su peso es mayor o igual que el promedio de los pesos”

En general, para determinar las entidades y relaciones relevantes se sigue el proceso explicado en la sección 4.4.4. Para determinar el peso de las Entidades se realiza la siguiente consulta en SPARQL:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX j.0: <http://www.semanticweb.org/MODS/ArbolAprendizaje#>

SELECT (AVG(?peso) AS ?promPeso) WHERE { { ?subject j.0:TieneTipo j.0:Nombre_Propio_re
. ?subject j.0:Peso ?peso} UNION { ?subject j.0:TieneTipo j.0:Nombre_Comun_re . ?subject
j.0:Peso ?peso}}
```

De igual manera, se realiza una consulta parecida para determinar el peso de las relaciones.

Una vez encontrado el promedio de los pesos de las entidades y relaciones, se determinan las entidades y relaciones relevantes. El GA es actualizado con dicha información.

Una vez obtenido el GA, se comienza con el proceso de actualización de los componentes ontológicos del MODS. Recordemos que existen tres casos de actualización, que se describieron en el capítulo 4, en la sección 4.4.5: 1) Actualización del Lexicón, 2) Actualización de Onomasticon, y 3) Actualización de la Ontología Interpretativa. En específico, el proceso de actualización de las ontologías del MODS, para nuestro caso de estudio, se muestra a continuación:

1. Actualización del Lexicón

Para extraer los sustantivos y los verbos, se utiliza la regla definida para ello en la sección 4.4.5 (es una consulta SPARQL sobre el GA). Una vez obtenido los sustantivos y verbos del GA, se procede a actualizar el lexicón, quedando el lexicón tal como se muestra en la Figura. 5.10 (los términos en los círculos rojos, son los nuevos términos incorporados al lexicón).

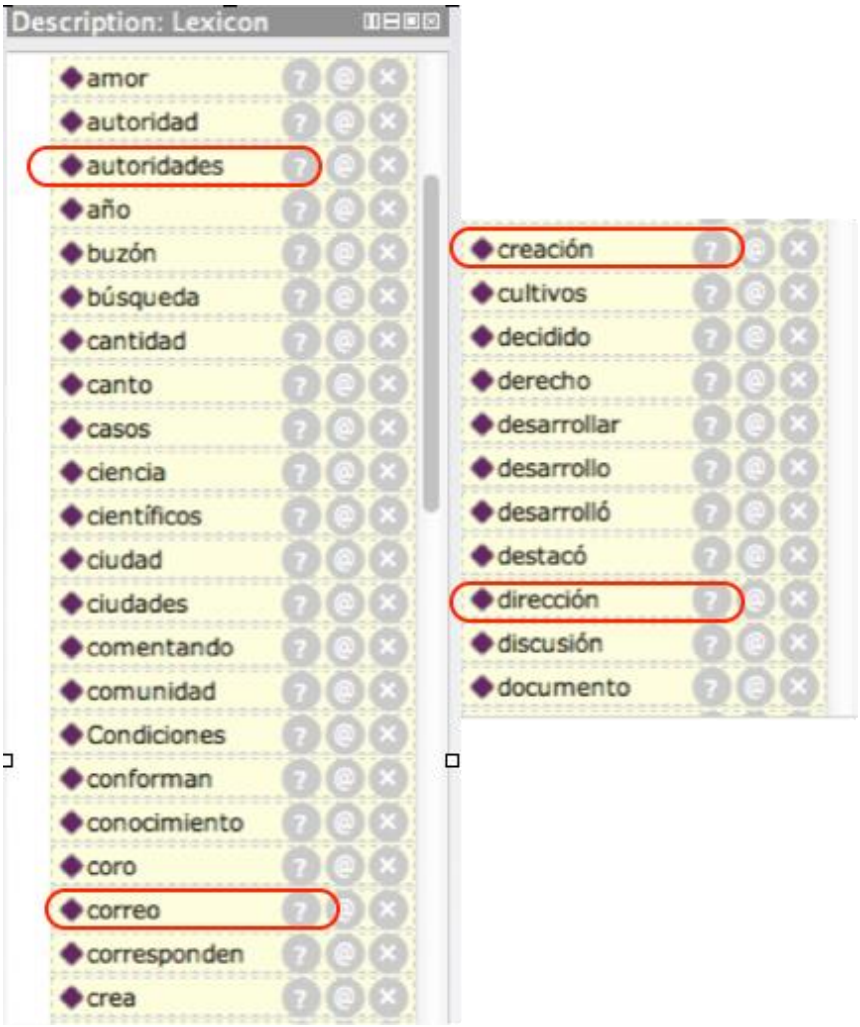


Figura 5.10. Ejemplo de parte del Lexicón, después de la actualización.

2. Actualización del Onomasticon

De igual manera que el lexicón, se procede actualizar el onomasticon, para la cual se aplica la regla definida en la sección 4.4.5 para extraer los nombres propios de la ontología semántica (GA) (es otra consulta SPARQL sobre el GA). Una vez obtenido los nombres propios del GA, se procede a actualizar el onomasticon, quedando el onomasticon, después de la actualización, como se muestra en la Figura. 5.11 (los nuevos términos incorporados son los que están en círculos rojos).

Figura 5.11. Ejemplo parte del onomasticon, después de la actualización.

3. Actualización de la Ontología Interpretativa

Para la actualización de la ontología interpretativa se sigue el procedimiento descrito en la Figura. 4.13 (ver capítulo 4). En la Figura. 5.12 se muestra la ontología interpretativa antes de la actualización.

Figura. 5.12. Ejemplo parte de la ontología Interpretativa antes de la actualización.

En la Figura. 5.13 se muestra un fragmento de la ontología semántica (GA) generada por el caso de estudio, donde se ven algunos de sus conceptos, relaciones taxonómicas y no taxonómicas.

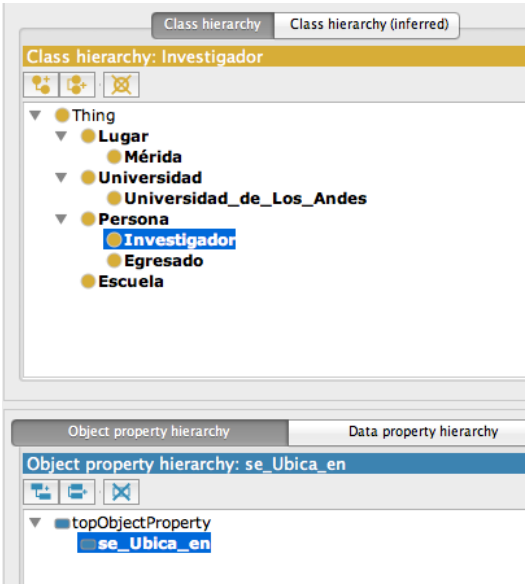


Figura. 5.13. Ejemplo parcial de la ontología semántica generada (GA) por el caso de estudio**.**

Después del proceso de mezcla de las dos ontologías (ontología interpretativa y GA), la ontología interpretativa es actualizada con las entidades, relaciones taxonómicas y no taxonómicas de GA, quedando como se muestra en la Figura. 5.14.

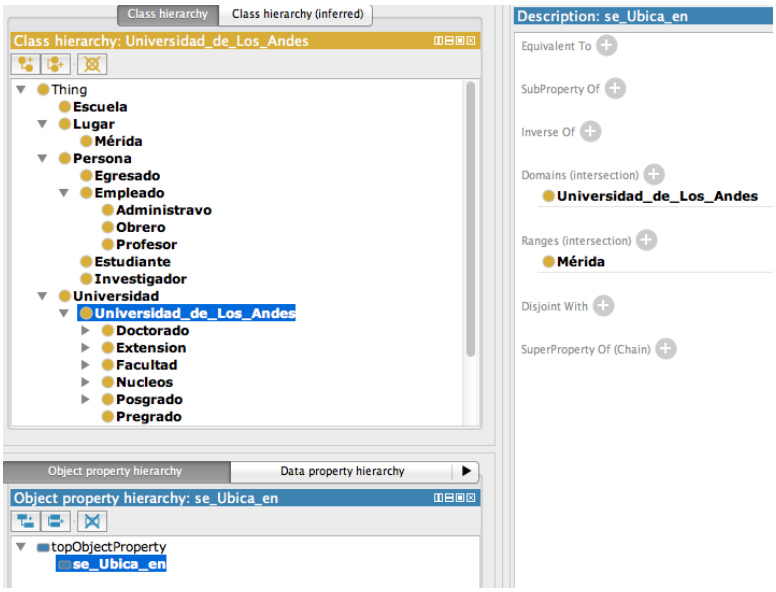


Figura. 5.14. Ejemplo parcial de la ontología interpretativa, después del proceso de actualización.

En la Figura. 5.14 se puede ver que se actualizaron *los conceptos*: Escuela, Lugar, Mérida, Investigador. Además, se actualizaron *las relaciones taxonómicas*: Universidad de Los Andes es una Universidad, Investigador es una Persona. Finalmente, se actualizó *la relación no taxonómica*: la Universidad de Los Andes se ubica en Mérida.

5.5. Minería Semántica

Como a grosos modo se mostró en la corrida anterior, para el desarrollo de los componentes de aprendizaje, se requieren del desarrollo de ciertos módulos/servicios de minería semántica, que básicamente permiten:

- ✓ Extraer documentos relevantes desde la Web
- ✓ Extraer conocimiento desde documentos no estructurados
- ✓ Generar Ontologías desde cero.
- ✓ Generar lexicones computacionales
- ✓ Generar Bases de datos terminológicas

En esta parte, vamos a analizar más en detalle las potenciales aplicaciones del sistema de extracción de conocimiento desde documentos no estructurados, en particular, del modelo semántico de conocimiento que subyace en él (la ontología semántica (GA) que se genera).

En específico, vamos a analizar su capacidad para generar la base terminológica, el lexicón electrónico, y la taxonomía es_un o es_una, del Doctorado en Ciencias Aplicadas de la Universidad de Los Andes, dinámicamente. Para ello, vamos a usar como base el sitio Web del doctorado (http://www.ing.ula.ve/doctorado_ca/). En este caso, se procesaron 12 documentos del sitio Web. Además, a esos documentos se les calculo las métricas definidas en el capítulo 4 sobre la frecuencia, la frecuencia inversa, y el peso (w) de los términos presentes en ellos, para determinar la relevancia de los mismos. Para dichos cálculos, se realizaron las siguientes consultas en SPARQL sobre la información recopilada de los documentos:

- Para determinar las entidades candidatas y el promedio de pesos de cada entidad, la consulta en SPARQL es la siguiente:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX j.0: <http://www.semanticweb.org/MODS/ArbolAprendizaje#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT (COUNT( ?subject) AS ?Entidades_Candidatas ) (AVG(?object) AS ?Promedio_Peso)
WHERE { {?subject j.0:Peso ?object . ?subject j.0:TieneTipo  j.0:Nombre_Propio_re } UNION { ?subject j.0:Peso ?object . ?subject j.0:TieneTipo  j.0:Nombre_Comun_re } }
```

- Para determinar las relaciones candidatas y el promedio de los pesos de cada relación, la consulta en SPARQL es la siguiente:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX j.0: <http://www.semanticweb.org/MODS/ArbolAprendizaje#>
SELECT DISTINCT (COUNT( ?Relación) AS ?Relaciones_Candidatas ) (AVG(?Peso) AS ?promPeso)
WHERE { ?Relación j.0:Peso ?Peso . ?Relación j.0:esUn j.0:Verbo_re }
```

Se obtienen los resultados mostrados en la Tabla 5.3. En esa tabla se puede observar que hay 602 entidades candidatas, y el promedio de pesos de las entidades es 4,89. Además, hay 205 relaciones candidatas, y el promedio de los pesos de las relaciones es 9,92.

Tabla. 5.3. Resumen de las entidades candidatas y relaciones candidatas

	Total	Promedio Peso
Entidades Candidatas	602	4,89
Relaciones Candidatas	205	4,92

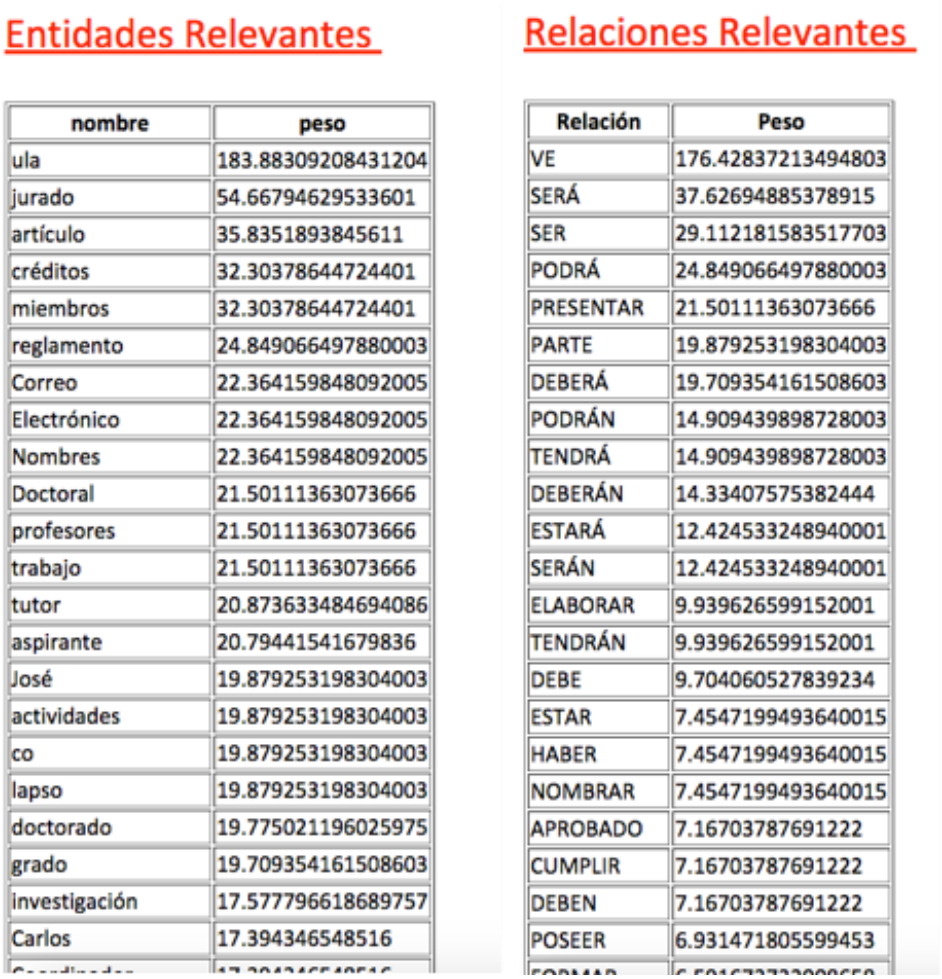
A partir de esos valores, se pueden determinar las entidades y relaciones relevantes, con las que se construirá el GA para este caso de estudio. La figura 5.15 muestra las entidades y relaciones relevantes, basada en el criterio utilizado para las selección de Entidades y Relaciones relevantes definidas en el capítulo 4: “todos los términos mayores o iguales que el promedio de los pesos”. La consulta específica en SPARQL para obtener las entidades relevantes, con sus respectivos pesos, desde la información recopilada de los documentos, es:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX j.0: <http://www.semanticweb.org/MODS/ArbolAprendizaje#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
SELECT DISTINCT ?nombre ?peso WHERE { {?subject j.0:tieneFrecuencia ?object . ?subject j.0:TieneTipo j.0:Nombre_Propio_re . ?subject j.0:tieneNombre ?nombre . ?subject j.0:TieneTipo ?tipo . ?subject j.0:Peso ?peso } UNION { ?subject j.0:tieneFrecuencia ?object . ?subject j.0:TieneTipo j.0:Nombre_Comun_re . ?subject j.0:tieneNombre ?nombre . ?subject j.0:TieneTipo ?tipo . ?subject j.0:Peso ?peso } FILTER (?peso>4.89)}
```

Con esta consulta se encuentra las entidades relevantes, las cuales son los sustantivos de “Nombres Propios” y “Nombres Comunes”.

El mismo proceso se sigue para determinar las relaciones relevantes. Siguiendo el mismo proceso, usando el valor respectivo de la Tabla 5.3 (pesos mayores o iguales a 4,92), y haciendo la consulta requerida en SPARQL sobre la información

recopilada de los documentos, se obtiene los resultados mostrados en la Figura. 5.15.



Entidades Relevantes

nombre	peso
ula	183.88309208431204
jurado	54.66794629533601
artículo	35.8351893845611
créditos	32.30378644724401
miembros	32.30378644724401
reglamento	24.849066497880003
Correo	22.364159848092005
Electrónico	22.364159848092005
Nombres	22.364159848092005
Doctoral	21.50111363073666
profesores	21.50111363073666
trabajo	21.50111363073666
tutor	20.873633484694086
aspirante	20.79441541679836
José	19.879253198304003
actividades	19.879253198304003
co	19.879253198304003
lapso	19.879253198304003
doctorado	19.775021196025975
grado	19.709354161508603
investigación	17.577796618689757
Carlos	17.394346548516

Relaciones Relevantes

Relación	Peso
VE	176.42837213494803
SERÁ	37.62694885378915
SER	29.112181583517703
PODRÁ	24.849066497880003
PRESENTAR	21.50111363073666
PARTE	19.879253198304003
DEBERÁ	19.709354161508603
PODRÁN	14.909439898728003
TENDRÁ	14.909439898728003
DEBERÁN	14.33407575382444
ESTARÁ	12.424533248940001
SERÁN	12.424533248940001
ELABORAR	9.939626599152001
TENDRÁN	9.939626599152001
DEBE	9.704060527839234
ESTAR	7.4547199493640015
HABER	7.4547199493640015
NOMBRAR	7.4547199493640015
APROBADO	7.16703787691222
CUMPLIR	7.16703787691222
DEBEN	7.16703787691222
POSEER	6.931471805599453

Figura 5.15. Entidades y Relaciones Relevantes

Por lo tanto, el GA que se genera en este caso está compuesto por esas entidades y relaciones relevantes (serán sus instancias). A partir de ese GA son muchas las cosas que se pueden hacer, ya que es una ontología semántica con todo el conocimiento extraído desde los documentos. A continuación vamos a describir algunas de ellas para este caso de estudio.

Partiendo del hecho de que GA contiene un conjunto de axiomas básicos que permiten inferir nuevo conocimiento, algunas de las potenciales aplicaciones usando un motor de inferencia sobre GA son:

- ✓ Construir una Base Terminológica en un dominio especializado.
- ✓ Crear lexicones electrónicos con los sustantivos y verbos descubiertos.
- ✓ Detectar la relación es_un (superclase-subclase), la cual se basa en la semejanza de los conceptos (uno de los conceptos incluye al otro), con la cual es posible construir taxonomías.

A continuación vamos a describir esos usos en este caso de estudio.

5.5.1. Base Terminológica del doctorado de ciencias aplicadas

GA contiene la lista de términos especializados de un dominio a partir del corpus de los documentos estudiados. Para ello se consideró el criterio de peso definido en el capítulo 4 para determinar los términos relevantes. Para ello, se podrían usar las siguientes reglas sobre GA:

TieneTipo(?x, Nombre_Propio_re) -> Onomasticon(?x),
TieneTipo(?x, Nombre_Comun_re) -> Lexicon(?x)
Onomasticon((?x)or(Lexicon((?x)-> TerminosEspecializados

Así, ese criterio permite caracterizar los términos (nombres propios, nombres comunes, etc.) relevantes (peso 4,89) que caracterizan al doctorado. Ese conjunto de términos especializados del doctorado contenido en el GA son mostrados en la Figura. 5.16.

Figura 5.16. Términos Especializados del Doctorado de Ciencias Aplicadas

Esta capacidad de adquisición especializada terminológica, se puede usar para el procesamiento textual, en tareas de minería de texto, en creación de ontologías especializadas, etc.

5.5.2. Lexicón Electrónico

En el caso en que se estuviese diseñando un lexicón electrónico de sustantivos y verbos, se aplicarían las siguientes reglas sobre el GA:

TieneTipo(?x, Nombre_Comun_re), -> Lexicon(?x)
esUn(?x, Verbo_re) -> Evento(?x)
Lexicon(?x)orEvento(?x)->LexiconElectronico

Estas reglas permiten extraer los sustantivos y verbos guardados en el GA. Para el caso del doctorado, se obtendrían los sustantivos y verbos mostrado en la Figura. 5.17.

Figura 5.17. Lexicón electrónico de sustantivos y verbos

En este caso, la posibilidad de construir y actualizar lexicones computacionales automáticamente (aprendizaje) abre un mundo a nivel lingüístico.

5.5.3. Creación de taxonomía es_un o es_una

En los textos analizados del Doctorado se encontraron varias sentencias que tienen el verbo es, como por ejemplo:

1. El doctorando es inmerso totalmente en la dinámica del grupo de investigación al cual pertenece su tutor, y sigue los lineamientos previamente establecidos por éste en el plan de formación.
2. Cualquier investigador cualificado que sea miembro de un grupo de investigación consolidado de la Universidad de Los Andes es, potencialmente, un tutor del programa.

3. Si el Plan de Formación no es aceptado por la Comisión de Admisión, el aspirante con su tutor podrán modificarlo y someterlo una vez más a la consideración de la Comisión, en un lapso de un mes.

Pasemos a describir como se usaría el GA para este caso. Antes que nada, se busca en que parte de los textos analizados del Doctorado se encuentra el verbo *es*, y se extraen las sentencias donde se encuentra, Para cada sentencia, se realiza todo el proceso de análisis de sentencias, para descubrir el uso del verbo *es* en la misma (ver sección 4.4.3). En particular, se busca determinar si en ellas existen relaciones candidatas del tipo "es un". Por ejemplo, para las tres sentencias arriba definidas, el resultado sería:

- En la sentencia 1 no se puede establecer ninguna relación "es un" entre entidades.
- En la sentencia 2 se establece la relación *investigador es un tutor*.
- En la sentencia 3 no se puede establecer la relación "es un" debido a que el verbo es "es aceptado".

Por lo tanto, para el caso anterior solo se podría obtener la relación taxonómica mostrada en la Figura. 5.18, la cual se almacenaría en GA.

Figura 5.18. Relación taxonómica de las sentencias analizadas.

En general, este proceso habría que hacerlo con todas las sentencias de todos los documentos analizados del doctorado. Esto alimentaria a GA con todas las relaciones taxonómicas existentes en los documentos.

Otra forma de usar el GA en este caso es determinar si en las relaciones relevantes contenidas en el GA aparece la relación "ser un" (es un evento). Si fuese el caso, cuando sean entidades las vinculadas por esa relación se estarían definiendo una relación taxonómica, y en caso contrario (cuando son nombres propios) serían instancias (individuos) de entidades ya existentes de esa *relación taxonómica* en el GA. Por ejemplo, si GA tuviera la relación relevante *es un* entre las entidades relevantes *investigador* y *tutor,* se podría determinar la relación taxonomía entre ellas dos.

De esta manera podemos crear automáticamente taxonomías "es_un", las cuales son las más comúnmente usadas al construir ontologías, y las mejor soportadas por las herramientas de desarrollo de ontologías (como protégé).

Vemos así estas tres formas de generación de conocimiento, a partir del GA que fue generado por el conjunto de documentos no estructurados del doctorado.

5.6. Análisis de resultados

En esta sección se van a realizar otro conjunto de pruebas al componente de aprendizaje.

5.6.1. Aprendizaje morfosintáctico

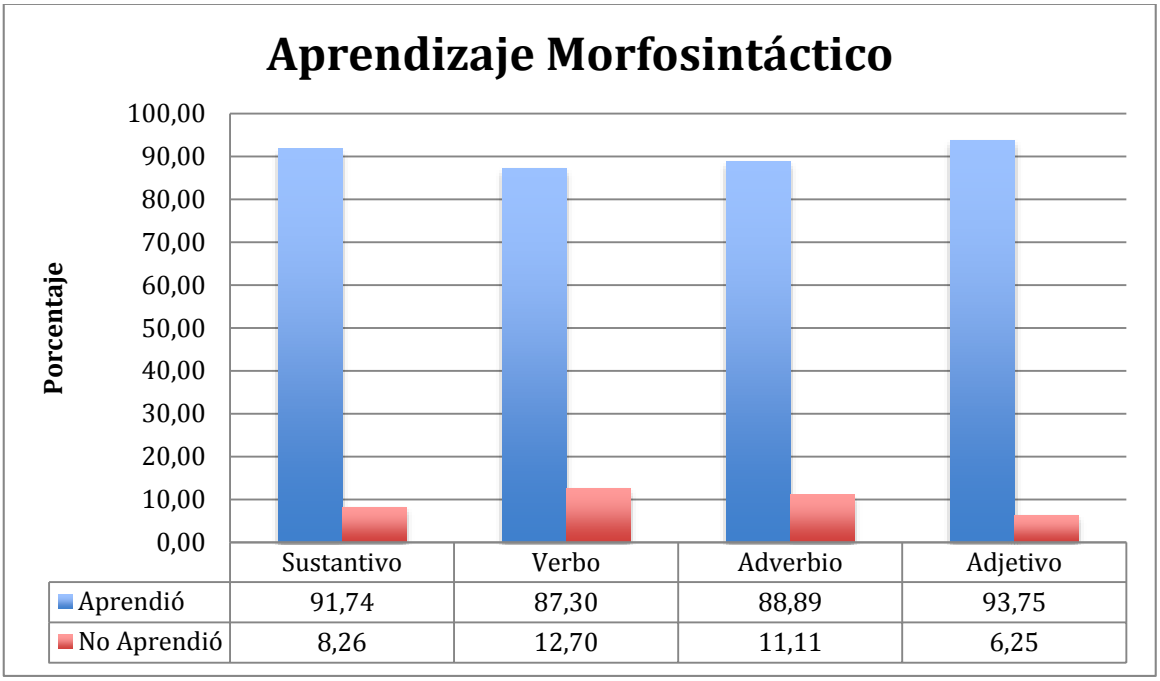
Las pruebas que se describen a continuación son para el caso de que el componente léxico no se encuentre en el lexicón. En este caso se quieren evaluar las capacidades de aprendizaje de nuevos léxico de nuestro módulo de aprendizaje morfosintáctico.

Para la prueba se introdujeron en el lexicón los verbos ser, estar, y algunos artículos y pronombres; luego se realizaron 500 consultas para demostrar que el lexicón va aprendiendo los términos desconocidos, tales como sustantivos, verbos, adjetivos y adverbios, encontrados en la consulta.

En la Figura 5.19 y en la Tabla 5.4 se muestran los resultados del aprendizaje morfosintáctico, en lo que se evidencia que del total de palabras por aprender, en total se aprendió en promedio más del 90%.

Tabla. 5.4. Resultados del proceso de aprendizaje de términos del aprendizaje morfosintáctico

Término	Aprendió		No Aprendió		Total
Sustantivo	500	91,74%	45	8,26%	545
Verbo	220	83,30%	32	12,70%	252
Adverbio	80	88,89%	10	11,11%	90
Adjetivo	75	93,75%	5	6,25%	80



	Sustantivo	Verbo	Adverbio	Adjetivo
Aprendió	91,74	87,30	88,89	93,75
No Aprendió	8,26	12,70	11,11	6,25

Figura. 5.19. Aprendizaje Morfosintáctico

Como se puede observar en los resultados, el porcentaje de aprendizaje de sustantivo es 91,74 %, lo cual era de esperarse, ya que los sustantivos son nombres propios y acrónimos, los cuales no necesariamente se encuentran en diccionarios electrónicos tales como WordReference.com y la real academia (en nuestros prototipo, el aprendizaje ágil se basa en la consulta a esos dos diccionarios). En cuanto a los verbos, el porcentaje es de 87,30% verbos aprendidos. Los verbos no aprendidos corresponden a verbos de propósito general, que en la implementación del algoritmo de aprendizaje de verbos no se contemplaron sus aprendizajes. Esos tipos de verbos son (traer, valer, salir, tener, venir, poner, hacer, decir, poder, querer, saber, andar, conducir), o el auxiliar (haber), o los copulativos (ser, estar), o los monosílabos (dar, ver, ir). Estos verbos, por sus particularidades (de irregularidad y número reducido), son más prácticos cargarlos manualmente en el lexicón, que diseñar algoritmos para su aprendizaje. En cuanto a los adverbios, 88,89% de ellos fueron aprendidos, la razón es que en la implantación no se contemplaron los adverbios que tienen sufijo ”mente”, ya que en los diccionario electrónicos usados por el aprendizaje ágil no se encuentra este tipo de adverbios. Finalmente, 93,75% de adjetivos son aprendidos, es la categoría lingüística que más se aprende debido a que el recurso que se seleccionó (WordReference.com) tiene casi la mayoría de los adjetivos existentes en su base de datos.

5.6.2. Aprendizaje semántico

En esta sección se procede a analizar, por medio de un ejemplo, la calidad del módulo de aprendizaje semántico del MODS.

Como se dijo en la sección 4.4 (ver también sección 5.4), la entrada del componente de aprendizaje semántico es un conjunto de enlaces Web. Para la misma consulta de la sección 5.4, se analizan los primeros 173 documentos recuperados, tanto por Yahoo como por Google. En general, se obtuvieron los resultados mostrados en la Tabla. 5.5. Por un lado, se encontraron 44 enlaces duplicados en la lista de resultados entre Yahoo y Google. Además, aplicando el primer criterio de filtrado inicial en la lista de resultados (archivos pdf y con autenticación de acceso), para los 173 documentos recuperados, se obtiene que 32 de Yahoo y 47 de Google son enlaces no relevantes, quedando 141 documentos relevantes candidatos de Yahoo y 126 de Google.

Tabla 5.5. Resultados del primer filtrado

Filtro	Yahoo	Google
Enlaces pdf	9	23
Enlaces con autentificación de acceso	23	24
Total de enlaces no relevantes	32	47
Total de enlaces relevantes candidatos en el primer filtro	141	126

Siguiendo con el segundo filtrado definido en la sección 4.4, utilizando los criterios tales como tipo de documento html, el lenguaje español y los enlaces inactivos, nos arroja una nueva tabla de documentos candidatos (ver Tabla. 5.6). Este filtrado implica ir a cada enlace, y si el enlace está activo extraer la información del tipo de documento y el lenguaje del documento. En este caso quedan 47 documentos candidatos en Yahoo y 105 en Google, para pasar a la fase del cálculo de la similitud entre la consulta y cada documento recuperado candidato. Así, con esos documentos se construye la matriz de frecuencia de los términos de la consulta en los documentos, para el cálculo de la similitud.

Tabla. 5.6. Resultados del segundo filtrado

Filtro	Yahoo	Google
Enlances pdf	9	23
Enlaces con autentificación de acceso	23	24
Total de enlaces no relevantes	32	47
Total de enlaces relevantes candidatos en el segundo filtro	141	126

Vemos que hasta aquí aún no se ha analizado la relación de similitud entre la consulta y los documentos candidatos. Sin aun usar ese criterio, que nos filtra si los resultados obtenidos cumplen con la necesidad de información establecida en la consulta, podemos ya calcular una primera medida de la precisión de la respuesta de los buscadores para los primeros criterios usados para determinar los documentos candidatos, usando la definición de esa medida hecha en el capítulo 2. La precisión define cuantitativamente la relación entre los documentos recuperados y su relevancia para satisfacer la consulta del usuario (ver Tabla. 5.7).

Tabla 5.7. Resultados de la medida de precisión

Filtro	Yahoo	Google
Total de Enlaces	173	173
Total de enlaces relevantes	47	105
Precisión	0,2716	0,6069

Según la precisión, se observa que de los criterios iniciales definidos por el usuario al realizar la consulta, solo el 60% de los documentos recuperados desde Google y el 27% de Yahoo los satisfacen. En particular, Google recupera más enlaces candidatos que Yahoo. Estos aún deben ser filtrados por el proceso de cálculo de la similitud (fase de determinación de documentos relevantes), para determinar si realmente esos documentos son útiles para el usuario (relevantes), lo que quizás hará aún más bajo la medida de precisión de cada buscador.

Es después de obtener los documentos relevantes que se procede a las actualizaciones de las ontologías del MODS, como se muestra a continuación: En general, para el ejemplo de la sección 5.4 se obtuvieron los siguientes resultados usando el MODS.

Tabla 5.8. Resultados de entidades y relaciones obtenidas para el ejemplo 5.4

	Candidatas	Relevantes
Entidades	2279	621
Relaciones	534	120
Total	2813	741

Actualización del lexicón

Con el GA se pudieron obtener nuevos sustantivos y verbos que no se encontraban en el lexicón. De las 621 Entidades relevantes existentes en el GA, se extrajeron los sustantivos del tipo Nombre_Comun_re, para lo cual se aplicaron las consultas SPARQL o reglas (axiomas) específicas sobre GA. En este caso se obtiene 256 sustantivos que se pueden agregar en el lexicón. En el momento de la actualización del lexicón, se verifica si el sustantivo está en el lexicón o no. En nuestro ejemplo se actualizó el lexicón con 102 sustantivos nuevos. De igual manera, se procede con la actualización de los verbos. En nuestro ejemplo se obtuvieron 120 verbos, y se actualizó con 48 verbos nuevos el lexicón, porque el resto ya estaba en el lexicón.

Actualización del Onomasticon

Para este proceso de actualización se sigue el mismo proceso que el del lexicón, pero en este caso se extraen los sustantivos del tipo Nombre_Propio_re. En nuestro ejemplo se encontraron 365 sustantivos de este tipo, y se actualizó el Onomasticon con 146 nuevos, ya que el resto estaba en el onomasticon.

Actualización de la Ontología Interpretativa

Para la actualización de la Ontología Interpretativa se realiza el proceso de mezcla entre GA y dicha ontología, tal como se explicó en la sección 5.4. Se parte de la hipótesis de que todos los sustantivos de tipo Nombre_Comun_re son conceptos, y todos los sustantivos de Nombre_Propio_re son individuos (instancias). Por lo tanto, los 256 sustantivos se convierten en conceptos en la nueva ontología mezclada. Una vez extraídos los conceptos, se comienza el proceso de extracción de las relaciones taxonómicas y no taxonómicas. En el caso de las relaciones taxonómicas se encontraron 10 nuevas, y en el caso de las relaciones no taxonómicas se encontraron 5 nuevas, con las cuales se actualizó la ontología interpretativa.

5.7. Análisis comparativo

En esta sección se hace un análisis cualitativo y comparativo del MODS con los trabajos previos (Ver Tabla. 5.9).

Tabla 5.9. Análisis cualitativo y comparativo

Sistema	Procesamiento de Lenguaje Natural	Formato de la Consulta	Recursos Lingüísticos	Lenguaje de la Ontología	Actualización automática de la Ontología
MESIA[70]	Propio	Booleana	ARIES y EWN	XML	No
QUACID[71]	Propio	SPARQL		RDF	No
PowerAqua[72]	Propio	SPARQL	Propio	RDF	No
SWSNL[73]	Propio	SPARQL	Propio	RDF, RDFS, OWL	No
MODS	Propio	Booleana SPARQL	Propio, corpus y diccionarios online	RDF, RDFS OWL	Si

En la tabla se puede observar que el MODS se diferencia de los trabajos previos en su proceso de actualización de las ontologías, el cual se realiza en forma automática. Además, para el procesamiento del lenguaje natural cuenta con la ontología de tarea, que describe cada una de las fases de análisis que se llevan a cabo (es una ontología que enlaza al resto de las ontologías del MODS, para el proceso de procesamiento de las consultas en lenguaje natural). Durante el procesamiento, como el resto de sistema, usa consultas SPARQL para obtener la información desde las ontologías del MODS. De los trabajos previos, el que más se asemeja al MODS es SWSNL, el cual tiene como principal diferencia que en el procesamiento del lenguaje natural no realiza el análisis sintáctico, solo realiza el análisis semántico, y no cuenta con el componente de aprendizaje.

[70] Modelo Computacional para Extracción Selectiva de Información de textos cortos
[71] Sistema de búsqueda de respuestas basado en ontologías, implicación textual y entornos reales.
[72] http://poweraqua.open.ac.uk:8080/poweraqualinked/jsp/index.jsp
[73] SWSNL: Semantic Web Search Using Natural Language

5.8. Conclusiones

En este capítulo se describió el funcionamiento del MODS, utilizando casos de usos, los cuales se explicaron en detalle en este capítulo. Cada caso de uso permitió verificar el funcionamiento de cada uno de los componentes del MODS, en especial el componente semántico, que es la principal contribución para la adaptabilidad del MODS.

Capítulo 6: Conclusiones

En este trabajo se propone un Marco Ontológico Dinámico Semántico (MODS) para la Web Semántica, con el fin de permitir interpretar y formalizar las consultas realizadas por los usuarios en lenguaje natural. Para ello, el MODS transforma la consulta en un formato de Representación del Significado (en nuestro caso, usa un hibrido para esa representación, basada en expresiones booleanas y en OWL, y consultas SPARQL), utilizando los diferentes componentes que lo componen para ese proceso de transformación: las ontologías del lexicón, onomasticon, tareas, lingüística e interpretativa. Además, MODS utiliza mecanismos de la minería semántica ontológica, y herramientas del Procesamiento de Lenguaje Natural, para lograr su objetivo.

Como se indicó antes, MODS cuenta con un conjunto de ontologías, y en particular, el proceso de integración de las ontologías la realiza la ontología de tareas para el procesamiento de la consulta en lenguaje natural. Dicha ontología utiliza los otros tres niveles de conocimiento (lexicón, lingüístico e interpretativo) del MODS, por lo que sirve de integradora para generar la consulta del usuario en un lenguaje ontológico, la cual será la que se envía a la Web.

Así, la ontología de tarea es una especie de meta-ontología que enlaza al resto de ontologías del MODS. Cada nivel de conocimiento guarda información muy específica requerida para el procesamiento del lenguaje natural, con sus propias dinámicas y autonomía funcional, que se integran en el momento adecuado del proceso, según lo vaya requiriendo la ontología de tareas.

Como parte de esa autonomía funcional, cada nivel de conocimiento tiene un mecanismo adaptativo, el cual forma parte del componente de aprendizaje del MODS. Dicho componente es el encargado de la adquisición automática de conocimiento, tanto léxico, morfológico como semántico (conceptos taxonómicos, no taxonómicos, relaciones, reglas de producción, etc.).

Para el mecanismo adaptivo, se propone una arquitectura para el aprendizaje automático de ontologías, capaz de adquirir nuevo conocimiento en función de la información suministrada por el proceso de análisis de una consulta en lenguaje natural, y/o de la información recuperada desde la Web por dicha consulta. Esta arquitectura, a partir del nuevo conocimiento adquirido, permite potenciar las estructuras ontológicas del MODS. La arquitectura de aprendizaje de ontologías se caracteriza por integrar diferentes métodos y técnicas de descubrimiento de conocimiento léxico, morfológico, sintáctico, semántico y pragmático. Además, a través de la arquitectura es posible explotar cualquier recurso de información, sea una palabra, una página Web, una base de datos, una ontología, etc. Así, se ha diseñado una arquitectura novedosa, con capacidad para soportar una variedad de elementos de aprendizaje. Pero lograr su implementación completa ha mostrado ser un reto de gran complejidad. Una de las razones es que estamos frente a uno de los fenómenos más interesantes en la inteligencia humana, el aprendizaje lingüístico.

En términos computacionales, esto implica la construcción de varios algoritmos que emulen los procesos que para los humanos son fáciles, como por ejemplo, saber si una palabra es un verbo o un sustantivo (por ejemplo: comer y pan). El caso anterior ocurre dentro de lo que se ha denominado *aprendizaje morfosintáctico*. Ese es uno de los casos estudiados en este trabajo. Otros casos de aprendizaje estudiados fueron, aprender conceptos, y relaciones taxonómicas y no taxonómicas los cuales forman parte de lo que se ha denominado *aprendizaje semántico*. A continuación se detallan los dos casos estudiados:

En cuanto al aprendizaje morfosintáctico, se diseñó e implementó un sistema de aprendizaje que, entre otras cosas, permite aprender verbos regulares e irregulares, y sus formas de conjugación en todas las formas personales e impersonales y en varios tiempos (aprendizaje duro). Además, tiene otros componentes (el aprendizaje simiente y el aprendizaje ágil), que le permiten aprender otras categorías lingüísticas: sustantivos, adjetivos y adverbios. En las pruebas realizadas vimos la importancia del aprendizaje duro para aprender verbos, y del aprendizaje ágil para aprender las otras categorías lingüísticas.

En cuanto al aprendizaje semántico, se realizó un sistema de aprendizaje de conceptos, relaciones taxonómicas y no taxonómicas, para documentos no estructurados que son recuperados de la Web, generando un grafo de aprendizaje (GA). Dicho GA es una ontología semántica, que en nuestro caso es representada en OWL, y cuyas instancias representan todo el conocimiento extraído de una colección de documentos. GA es posible usarlo para generar nuevo conocimiento semántico, como por ejemplo bases terminológicas especializadas, lexicones, taxonomías.

En particular, para la construcción del GA hemos usado una métrica que permite clasificar los términos contenidos en los documentos procesados como relevantes o no. Dicha métrica combina la frecuencia y la capacidad de discriminar de los términos contenidos en los documentos procesados. Esto le confiere al GA una característica fundamental, ser una ontología semántica de documentos procesados, conformada solamente por información (individuos) relevante.

Adicionalmente, hemos definido una serie de métricas, inspiradas en la minería de texto, para determinar cuando los documentos son relevantes para un usuario o no. Esto, al igual que el concepto de información relevante, permite caracterizar el GA. Por otro lado, el MODS usa el GA para actualizar a la ontología interpretativa, al onomasticon, y al lexicón. De esa manera, todos estos componentes del MODS se van adaptando a los perfiles de los usuarios.

En cuanto a los objetivos específicos definidos en este trabajo, los mismos los hemos cubiertos de la siguiente manera:

- ✓ *Profundizar en los aspectos teóricos relacionados a los temas de Web Semántica, Aprendizaje de Ontologías y Semántica Ontológica.* Este objetivo se alcanzó de dos manera, con la presentación del marco teórico base de esta investigación (ver capítulo 2), y con la revisión bibliográfica del estado de arte de cada uno de los temas relacionados a este trabajo (algunos presentados en el capítulo 2, y otros en el capítulo 4).

- ✓ *Describir formalmente las Ontologías utilizadas en el MODS.* Este objetivo se alcanzó con la propuesta de la arquitectura del Marco Ontológico Dinámico Semántico, y el diseño detallado de las ontologías lexicón, onomasticon, lingüística, interpretativa, y de tareas, usando para ello a OWL (todo esto se presentó en el capítulo 3).
- ✓ *Caracterizar el proceso de procesamiento de las consultas en lenguaje natural para la Web usando ontologías*. Este objetivo se alcanzó con el diseño en específico de la ontología de tarea, donde se integran todas las ontologías del MODS (realiza el enlazamiento del resto de ontologías de MODS), e invoca el componente de aprendizaje (ver capítulo 3).
- ✓ *Proponer modelos de aprendizaje automático de ontologías*. Este objetivo se alcanzó con la propuesta del componente de aprendizaje del MODS, y el desarrollo de los mecanismos/sistemas de aprendizaje morfosintáctico y semántico (todo eso está en el capítulo 4).
- ✓ *Probar el Marco Ontológico Dinámico Semántico para la Web Semántica en casos de estudio.* Este objetivo se alcanzó con el desarrollo de casos de estudios para cada componente, y para el MODS como un todo (ver el capítulo 5, que presenta todos los casos de estudios).

Finalmente, en cuanto a los trabajos analizados durante esta investigación, el principal aporte de MODS es su componente de aprendizaje, ya que en todos los estudiados carecen de este componente. Este componente tiene la ventaja que puede ser usado con el MODS, o de forma independiente para la construcción de ontologías, bases terminológicas sobre un dominio, lexicones computacionales, etc., todas ellas pudiéndose crear desde cero. Otro aporte importante es la ontología de tarea, la cual es una ontología integradora que orquesta el funcionamiento del MODS. A partir de ella se pueden agregar otros componentes que mejoren el funcionamiento del MODS, sin afectar los anteriores. Por último, es importante señalar que el lexicón, el onomasticon y la ontología lingüística, son ontologías que pueden ser usadas individualmente por cualquier sistema de procesamiento semántico.

En la versión actual se requieren algunos ajustes para mejorar el funcionamiento del MODS, uno de ellos es la adquisición automática de reglas lingüísticas para alimentar la ontología lingüística; otro ajuste es sobre el componente de aprendizaje, que actualmente está implementado solo para documentos no estructurados, por lo cual se requiere extender el componente para documentos estructurados. También, este componente de aprendizaje requiere de la intervención del ingeniero de conocimiento para la validación de los datos introducidos en la ontología interpretativa. En cuanto el tiempo de ejecución del componente de aprendizaje, depende de algunos factores, tales como la conexión de la red, las características de la máquina donde se ejecute, entre otros.

Entre los trabajos futuros se tienen los siguientes:

- ✓ *En cuanto a la ontología lingüística*, agregar al componente de aprendizaje un mecanismo para el descubrimiento de reglas lingüísticas (reglas sintácticas y

semánticas) en forma automática, al igual que se hace con los conceptos en la ontología interpretativa.

- ✓ *En cuanto la ontología de tareas*, se debe generalizar para ser usada en cualquier proceso de procesamiento semántico, ya que en los actuales momentos esta ontología está especializada para el Procesamiento del Lenguaje Natural. Adicionalmente, esta ontología puede ser aún más generalizada, para supervisar cualquier proceso de procesamiento del lenguaje (de texto, de audio, etc.).
- ✓ *El MODS* ha sido definido para la interpretación de consultas, pero se puede extender para la interpretación de textos, generación de resúmenes, usando las ontologías del MODS.
- ✓ *El componente de aprendizaje desarrollado* se utilizó en documentos no estructurados, pero se le pueden agregar mecanismos para analizar documentos heterogéneos (y en particular, documentos estructurados), tales como: Ontologías, Bases de datos, etc., para mejorar su proceso de aprendizaje.
- ✓ MODS tiene un conjunto de subsistemas (de extracción de documentos, de clasificación de sitios Web, etc.), los cuales podrían ser convertidos en servicios de minería semántica, para poder ser usados por aplicaciones que lo requieran.
- ✓ En general, cada uno de los componentes del MODS se puede convertir en servicios, para ser usado en cualquier sistema de procesamiento del lenguaje natural.
- ✓ Finalmente, algunas potenciales aplicaciones donde el MODS podría ser usado como base, con las extensiones respectivas, seria en traductores (se requeriría de los lexicones de los idiomas a traducir, y la ontología lingüística que hiciera el emparejamiento entre los idiomas), y en generación de resúmenes desde un grupo de documentos (se requeriría de un sistema generador de texto (y en este caso de resúmenes, para lo cual habría que definir un patrón de resúmenes a concebir), que use como insumo el GA).

Bibliografía

[1] Aguilar, J. “Sistema semántico para la búsqueda inteligente de información por contexto para la Web”, Revista Ciencia e Ingeniería, Vol. 32 Nro. 3, pp. 141-152, (2011)

[2] Albuín, N. y Vinader, R. “El desarrollo de la Word Wide Web en España: Una Aproximación Teórica desde sus orígenes hasta su transformación en un medio semántico”. Revista Razón y Palabra. Vol. 16. Nro 75. Febrero-Abril. (2011) Disponible en: http://wwww.razonypalabra.org.mx/N/N75/varia_75/varia3parte/31_Avuin_V75.pdf [Consultado 13 de julio de 2014]

[3] Antoniou, G. and Harmelen, F. “A Semantic Web Primer”. Second edition. The MIT Press. Cambridge, Massachusetts. London, England. 2008. ISBN 978-0-263-01242-3.

[4] Basterrechea, E. y Rello, L. “El verbo en español: Construye tu propio verbo” Primera edición 1.0. 2010. Molino de Ideas

[5] Berners-Lee, T., Hendler, J. y Lassila O., “The Semantic Web”. Scientific American, Mayo 2001.

[6] Berrocal, C. García, E. Rodríguez, A. Zazo. “Agentes inteligentes recuperación autómata de información en la Web”. Revista española de documentación científica. ISSN 0210-0614, Vol. 26, Nro 1, pp. 11-20. (2003). Disponible en: http://dialnet.unirioja.es/servlet/articulo?codigo=622775.

[7] Bláquez, M. “Técnicas avanzadas de recuperación de información. Procesos, técnicas y métodos”. Madrid:mblazquez.es. ISBN: 978-84-695-8030-1. Madrid. (2013)

[8] Bourigault, D. “LEXTER, un Logiciel d'EXtraction de TERminologie. Application à l'acquisition des connaissances à partir de texts”. Tesis Doctoral. Paris: École des Hautes Études en Sciences Sociales. (1994).

[9] Castelo, D., Isi, I., Martinez, S., Garat, G. and Moncecchi, Q. “WebQA: Respuesta Automática a Preguntas”. XXXIV Conferencia Latinoamericana de Informática (CLEI 2008). Disponible en: http://www.fing.edu.uy/inco/grupos/pln/publicaciones/PaperWebQA.pdf

[10] Chaelandar, G. y Grau, B. “SVETLAN’- A System to Classify Words in Context”. In S. Staab, A. Maedche, C. Nedellec, P. Wiemer-Hastings (eds.) Proceedings of the Workshop on Ontology Learning, 14th European Conference on Artificial Intelligence ECAI’00, Berlin, Germany, August 20-25. (2000).

[11] Fernández, O., Izquierdo, R., Fernández, S. y Vicedo, J. “Un sistema de búsqueda de respuestas basado en ontologías, implicación textual y entornos reales”. Procesamiento del lenguaje Natural. ISSN 1989-7553. Vol. 41, pp. 47-54. (2008), Disponible; http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/2548/1087

[12] Galicia S. Análisis sintáctico conducido por un diccionario de patrones de manejo sintáctico para lenguaje español. Instituto Politécnico Nacional. (2000). México, D.F. Disponible en: http://www.gelbukh.com/Tesis/Sofia/tesisfinal.htm

[13] Gelbukh, A. y Calvo, H. “Determinación automática de roles semánticos usando preferencias de selección sobre corpus muy grandes”. Computación y Sistemas. Vol 12, No 001, pp. 128-150. (2008). Disponible en: http://www.revistas.unam.mx/index.php/cys/issue/view/224/showToc

[14] Gómez A, Fernández M, Corcho O. “Ontological Engineering with examples from the areas of knowledge Management, e-Commerce and the Semantic Web”. ISBN:1-85233-551-3. Spinger-Verlaf London Limited (2004)

[15] Goñi, J. González, J. Moreno A. ARIES: A Lexical Plataform for Enginnering Spanish Processing Tools. Natural Language Engineering, Vol.3, Nro. 4; pp. 317-345. (1997) Disponible en: http://oa.upm.es/4739/

[16] Gruber, T. “A Translation Approach to Portable Ontology Specifications”. Edge Acquisition, 5(2), pp. 199-220, (1993).

[17] Guía Breve de Web Semántica. Disponible en: http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica

[18] Habernal, I. Konopík, M. “SWSNL: Semantic Web Search Using Natural Language”. Expert

Systems with Applications 40, pp 3649–3664, (2013)

[19] Hebeler, J. Matthew, F. Blace, R. Perez-Lopez, A. "Semantic Web Programming". by Wiley Publishing. ISBN:978-0-470-41801-7 (2009)

[20] Jacquin, C. y Liscouet, M."Terminology extraction from texts corpora: application to document keeping via Internet". In: TKE '96: Terminology and Knowledge Engineering, pp. 74-83. Berlin: IndeksVerlag. (1996)

[21] Justeson, J. and Katz, S. "Technical terminology: some linguistic properties and an algorithm for identification in text". Natural Language Engineering, Vol. 1, Nro. 1 pp. 9-27. (1995).

[22] Langley, P., "Elements of Machine Learning". Morgan Kaufmann Publishers, California, (1996).

[23] Lehmann, J. "DL-Learner: Learning Concepts in Description Logics" Journal of Machine Learning Research 10 pp. 2639-2642. (2009)

[24] Lopez, V., Sabou, M., Uren, V., and Motta, E. "Cross-Ontology Question Answering on the Semantic Web: an initial evaluation." Knowledge Capture Conference, pp. 17-24. Sept (2009), California. http://technologies.kmi.open.ac.uk/poweraqua/publications-downloads.php

[25] Maedche, A. y Volz, R. "The Text-To-Onto Ontology Extraction and Maintenance Environment". To appear in Proceedings of the ICDM Workshop on integrating data mining and knowledge management, San Jose, California, USA. (2001).

[26] Maedche, R., and Staab, S. "Ontology Learning for the Semantic Web". IEEE Intelligent Systems, pp 72-79, march-april (2001)

[27] Mahesh, K, and Nirenburg, S. "A Situated Ontology for Practical NLP". In Proceedings of Workshop on Basic Ontological Issues in egge Sharing. International Joint Conference on Artificial Intelligence, Montreal, Canada, August (1995)

[28] Manning, C, Raghavan, P and Hinrich, H., "Introduction to Information Retrieval" Cambrige, University Press. ISSB: 978-0-521-86571-5. 2008.

[29] Manual De Nociones Y Ejercicios Gramaticales Unidad De Composición Y Otras Destrezas Lingüísticas by Facultad De Estudios Generales Departamento De Español. ISBN 13: 9780847731640. ISBN 10: 0847731642

[30] Martí, M. "Tecnología del Lenguaje". Barcelona: Editorial UOC. pp. 285. (2003).

[31] Martínez, A. Segovia, F. "Introducción a la Ingeniería del software: Modelos de desarrollo de programas". ISBN: 84-96477-00-2. (2005).

[32] Martínez, J. "Los modelos clásicos de recuperación de información y su vigencia" Tercer Seminario Hispano-Mexicano de investigación en Bibliotecología y Documentación. UNAM, Centro Universitario de Investigaciones Bibliotecológicas. pp. 187-206. Disponible en: http://eprints.ucm.es/5979/1/Modelos_RI_preprint.pdf

[33] Martínez, P. y García, A. "Utilizando recursos lingüísticos para mejora de la recuperación de información en la Web". Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial. ISSN: 1137-3601. Vol 6, Nro. 016, pp. 55-66. (2002)

[34] Meadche A. and Staab, S. "Discovering conceptual relations from text". In W.Horn, editor, Proceedings of the 14th European Conference on Artificial Intellignece (ECAI'2000), 2000.

[35] Missikoff, M. Navigli, R. y Velardi P. " The Usable Ontology: An Environment for Building and Assessing a Domain Ontology". Research paper at International Semantic Web Conference (ISWC), Sardinia, Italia, June 9-12th (2002).

[36] Moreno N. "Semantica y Lexicologia de la Lengua Espanola". Disponible en: www.academia.edu

[37] Moreno, A. "Diseño e Implementación de un Lexicón Computacional para Lexicografía y Traducción Automática". Facultad de Filosofía y Letra. Universidad de Málaga. ISSN: 1139-8736. (2000).

[38] Morik, K. and Wroble, S. and Kietz, J. and Emde, W. "Knowledge acquisition and machine learning: Theory, methods, and applications", London: Academic Press, (1993).

[39] Nadeau, D. and Sekine, S., "A survey of named entity recognition and classification". En: Lingvisticae Investigationes 30-3-26,. – ISSN 0378–4169, Januar, Nro. 1.(2007)

[40] Nirenburg, S. y Raskin, V. (2004) Ontological Semantics (Draft) http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.375.8833&rep=rep1&type=pdf

[41] Onyshkevych. B and Nirenburg. S. “Lexicon, Ontology and Text Meaning”. Proceedings of the First SIGLEX Workshop on Lexical Semantics and Knowledge Representation. ISBN:3-540-55801-2 pp. 289-303. (1992).

[42] Puerto, E. Aguilar, J. Rodríguez, T. “Automatic Learning of Ontologies for the Semantic Web: experiment lexical learning”. Respuestas: Revista de la Universidad Francisco de Paula Santander. Año 17. No. 2, ISSN 0122-820X. pp. 5-12. diciembre (2012)

[43] Ramshaw, L. and Weischedel, R. “Information extraction”. En: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (2005)

[44] Rodríguez, Taniana y Aguilar, Jose. “Servicio para la Extracción de Documentos Relevantes desde Documentos Recuperados en la Web”. Centro de Estudios en Microelectrónica y Sistemas Distribuidos (CEMISID). Universidad de Los Andes. (ULA). Mérida-Venezuela. Enero (2015).

[45] Ruiz, A., Martínez, P, y García, A. “Análisis y expansión de consultas en lenguaje natural para mejora de la búsqueda en Web”. Procesamiento del Lenguaje Natural. ISSN 1989-7553. Vol. 27. Septiembre (2001). Disponible: http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/3379.

[46] Saiz, M. “Procesamiento del Lenguaje Natural: presente y perspectivas futuras”. Universidad de Alicante. España. (2003).

[47] Shamsfard, M. y Barforoush, A., “An introductionto Hasti: An Ontology learning System”.

[48] Willett, P. The Porter stemming algorithm: then and now. Program: electronic library and information systems, 40 (3). pp. 219-223. (2006)

[49] Wimalasuriya, D. and Dou, D. “Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches”. Journal of Information Science. Vol. 36, Nro. 3, pp-306-323. June (2010).

Anexo A. Relaciones Semánticas

Las relaciones semánticas: sinónimos [74] (synonymy), hipónimo [75] (hyponym), antónimos[76] (antonymy), Meronimias[77] (meronymy). Además, usa el estándar de codificación léxica EAGLES [78] y su clasificación semántica definida por: organización, lugar, persona y otro.

Tabla A.1. Relaciones semánticas del Lexicón

Relación Semántica	Patrón Lenguaje Natural	Relación
Sinónimos	• X "puede ser también llamado" Y • X "es también llamado" Y • X ", también llamado" Y • X "sinónimo de" Y • X "denominado/s" Y • X "ser un" Y • X "equivalente de" Y • X "ser igual que" Y • X "ser equivalente a" Y • X "ser un equivalente de" Y • X "ser lo mismo que" Y • X "ser idéntico a" Y • X "equivaler a" Y • X "ser sinónimo de" Y • X "ser un sinónimo de" Y • X "llamado/s" Y	Sinonimo (X, Y) La relación de sinónimo es una relación binaria Propiedad ser: Simétrica $Si\ X\ Relacion\ Y \Rightarrow Y\ Relacion\ X,$ $entonces\ X\ Sinonimo\ de\ Y$
Hipónimo	• A "es una" B • A "es un" B • A "es un tipo de" B • A "de cierta manera es" B • A "en cierto sentido es" B	Hiponimo(A, B) $Si\ B \sqsubseteq B(Refexividad),$ $B \sqsubset A \Rightarrow \neg A \sqsubset B\ Antisemetría,$ $B \sqsubset A\ Z \sqsubset B \Rightarrow\ Z\ \sqsubset\ A(Transitividad),$ $entonces\ B\ es\ un\ hipónimo\ de\ A$
Antónimo	• X "opuesto de" Y	Antonimo(X,Y) $Si\ X\ Relacion\ Y \Rightarrow \neg Y\ Relacion\ X,$ $entonces\ X\ Antonimo\ de\ Y$
Meronimia	• X "estar formado por" Yi • X "estar compuesto por" Yi • X "formar parte de" Yi	Meronimia $Si\ X\ \sqsubset\ Y$ $Y_1 U Y_2 \ldots Y_n$ $donde\ Y_i \sqsubseteq \neg Y_j\ 1 \leq i \leq j \leq n$

[74] Sinónimo: Vocablos o expresiones que tienen una misma o muy parecida significación: "esposos" y "cónyuges" son términos sinónimos. http://www.wordreference.com/definicion/sinonimos

[75] Hipónimo: palabra cuyo significado es más específico que otro, y es englobado por ese otro significado: "minuto" es un hipónimo de "tiempo". http://www.wordreference.com/definicion/hip%C3%B3nimo

[76] antónimo: Palabra que expresa una idea opuesta o contraria a la expresada por otra palabra: "vicio" y "virtud" son palabras antónimas; el antónimo de "claro" es "oscuro".http://www.wordreference.com/definicion/ant%C3%B3nimo

[77] Meronimia: es la correspondencia léxica de la relación parte-todo. Decimos que el dedo es un merónimo de la mano [7].

[78] http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

Anexo B. Ontología lingüísticas

A continuación se presenta parte de las reglas sintácticas implementadas en el MODS

Tabla B.1. Reglas sintácticas

Reglas	Rule
1. FraseNominal FraseVerbal -> Oración	1. FraseNominal(?x), FraseVerbal(?y), hasNext(?x, ?y) -> Oracion(?z)
2. Verbo-> FraseVerbal	2. Verbo(?x), hasNext(?x, null) -> FraseVerbal(?x)
3. NombrePropio->FraseNombre	3. NombrePropio(?np), endFN(?np, null) -> FraseNombre(?np)
4. NombrePropio Preposicion NombrePropio-> FraseNominal	4. NombrePropio(?np), NombrePropio(?np1), Preposicion(?p), endFN(?np1, null), hasNext(?np, ?p), hasNext(?np1, null), hasNext(?p, ?np1), hasPrevious(?np, null) -> FraseNominal(?np)
5. NombreComun-> FraseNombre	5. NombreComun(?x), endFN(?x, null), hasNext(?x, null) -> FraseNombre(?x)
6. Verbo FraseNominal -> FraseVerbal	6. FraseNominal(?fn), Verbo(?x), hasNext(?x, ?fn) -> FraseVerbal(?x)
7. Verbo NombreComun -> FraseVerbal	7. NombreComun(?y), Verbo(?x), hasNext(?x, ?y), hasNext(?y, null) -> FraseVerbal(?x)
8. FraseNombre -> FraseNominal	8. FraseNombre(?np) -> FraseNominal(?np)
9. NombrePropio -> FraseNombre	9. NombrePropio(?np), endFN(?np, null) -> FraseNombre(?np)
10. FraseNominal -> Oracion	10. FraseNominal(?FN) -> Oracion(?FN)
11. FraseVerbal -> Oracion	11. FraseVerbal(?x) -> Oracion(?x) 12. Articulo(?x), NombreComun(?nc), endFN(?nc, null), hasNext(?x, ?nc), hasPrevious(?x, null) -> FraseNominal(?x)
12. Articulo NombreComun -> FraseNominal	13. Articulo(?x), NombreComun(?nc), endFN(?nc, null), hasNext(?x, ?nc), hasPrevious(?x, null) -> FraseNominal(?x)

Verbos_Copulativos tienen estructura Sintáctica como Semántica:

$$Verbos_{Copulativos} \subseteq Verbos \text{ ^ } tiene.EstructuraSintactica \text{^ } tiene.EstructuraSemantica$$

Para los sustantivos, adjetivos y adverbios ejemplos de axiomas son

$$Sustantivos \subseteq Sustantivos \text{ ^ } tiene.EstructuraSintactica \text{^ } tiene.EstructuraSemantica$$

$$Adjetivos \subseteq Adjetivos \text{ ^ } tiene.EstructuraSintactica \text{^ } tiene.EstructuraSemantica$$

$$Adverbios \subseteq Adverbios \text{ ^ } tiene.EstructuraSintactica \text{^ } tiene.EstructuraSemantica$$

En este trabajo solo nos vamos a centrar en los verbos transitivos

El cual la estructura sintáctica y semántica es la siguiente:

1. $EstructuraSemantica \subseteq Agente \cap Tema \cap Beneficiario$
2. $EstructuraSemantica \subseteq Agente \cap Tema$

Y las estructuras sintácticas de 1 y 2, son las siguientes:

1. $EstructuraSintactica \subseteq Sujeto \cap Objecto\_Directo \cap Objecto\_Indirecto$
2. $EstructuraSintactica \subseteq Sujeto \cap Objecto\_Directo$

En el capítulo 3, se presenta algunos ejemplos de la estructura semántica

A continuación se presenta algunos ejemplos de reconocimiento de la estructura semántica mostrada en Protégé

1. **juan Corre**
   Después de razonar con pellet, se tiene lo siguiente:

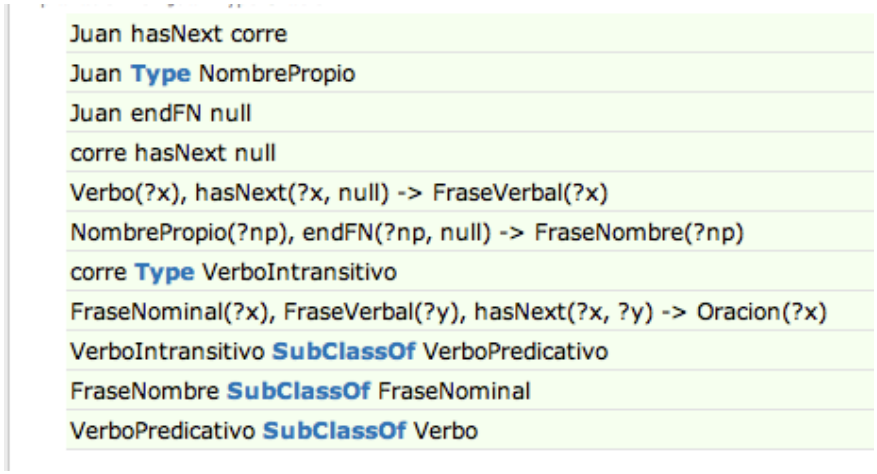


Figura. B.1. Análisis Sintáctico de “Juan Corre”

2. Facultad de Ingeniería

Después de razonar con pellet, se tiene lo siguiente

Facultad hasNext de
Facultad hasPrevious null
Facultad **Type** NombrePropio
NombrePropio(?np), NombrePropio(?np1), Preposicion(?p), endFN(?np1, null), hasNext(?np, ?p), hasPrevious(?np, null) -> FraseNominal(?np)
Ingeniería endFN null
de hasNext Ingeniería
FraseNominal(?FN) -> Oracion(?FN)
Ingeniería hasNext null
Ingeniería **Type** NombrePropio
de **Type** Preposicion

Figura B.2. Análisis Sintáctico de “Facultad de Ingeniería”

3. El velocista salta obstáculos

Después de razonar con pellet, se tiene lo siguiente:

el hasNext velocista
el **Type** Articulo
velocista **Type** NombreComun
velocista endFN null
Articulo(?x), NombreComun(?nc), endFN(?nc, null), hasNext(?x, ?nc) -> FraseNomin

salta **Type** VerboTransitivo
VerboTransitivo **SubClassOf** VerboPredicativo
VerboPredicativo **SubClassOf** Verbo
salta hasNext obstáculo
obstáculo hasNext null
NombreComun(?y), Verbo(?x), hasNext(?x, ?y), hasNext(?y, null) -> FraseVerbal(?x)
obstáculo **Type** NombreComun

1)	el_velocista **Type** FraseNominal
2)	el_velocista hasNext salta_ostaculos
3)	salta_ostaculos **Type** FraseVerbal
4)	FraseNominal(?x), FraseVerbal(?y), hasNext(?x, ?y) -> Oracion(?x)

Figura B.3. Análisis Sintáctico de “El velocista salta obstáculos”

Anexo C. Axiomas de la Ontología de Tareas

La representación de la Ontología de Tareas se realizó con lógica descriptiva, a continuación se muestran un ejemplo de los axiomas en su base de conocimiento:

$$AnalisisLexicoMorfologico \sqsubseteq Tarea \cap \forall Ejecuta_{Metodo}.AnalisisLexico \cap \forall Ejecuta_{Metodo}.AnalisisMorfologico$$

Este axioma indica que para que se ejecute la tarea de AnalisisLexicoMorfologico se deben ejecutar los métodos de Análisis Léxico y Morfológico. A continuación se muestra un ejemplo de las reglas que se derivan del axioma definido antes:

$$rule :: Implies($$

$$Antecedente(AnalisisLexico(E: Consulta?\, x, S: Vector_{Tokens}?\, y)\ \wedge$$

$$AnalisisMorfologico(E: Vector_{Tokens}?\, y, S: Vector_{Estructura\_Sintactica}\ ?\, z)$$

$$Consecuente\left(AnalisisLexicoMorfologico\left(\begin{matrix} E: Consulta?\, x, \\ S: Vector_{Estructura\_morfológicas}\ ?\, z \end{matrix}\right)\right)$$

)

Esta regla nos indica que se debe ejecutar el Análisis Léxico (recibe como entrada la consulta realizada por el usuario y da como salida un vector de tokens), y el Análisis Morfológico (recibe como entrada el vector de tokens y da como salida el Vector que contiene todas las posibles estructuras morfológicas de las palabras de la consulta). Eso es el análisis léxico-morfológico de MODS.

A continuación se muestra algunas de las reglas definidas en la ontología de tareas

	Reglas
1.	$AnalisisLexicoMorfologico \sqsubseteq Tarea \cap \forall Ejecuta_{Metodo}.AnalisisLexico \cap \forall Ejecuta_{Metodo}.AnalisisMorfologico \cap \forall Ejecuta_{Metodo}.ConsultaLexicon$
2.	$AnalisisLexicoMorfologico \sqsubseteq Tarea \cap \forall Ejecuta_{Metodo}.AprendizajeMorfosintáctico$
3.	$AnalisisSintactico \sqsubseteq Tarea \cap \forall Ejecuta_{Metodo}.ReglasSintacticas \cap \forall Ejecuta_{Metodo}.ConsultaOntologiaLinguistica \cap \forall Ejecuta_{Metodo}.CompararReglasSintacticas \cap \forall Ejecuta_{Metodo}.GenerarArbolSintactico$
4.	$AnalisisSintactico \sqsubseteq Tarea \cap \forall Ejecuta_{Metodo}.AprendizajeLinguistico$
5.	$AnalisisSemantico \sqsubseteq Tarea \cap \forall Ejecuta_{Metodo}.ReglasSemantica \cap \forall Ejecuta_{Metodo}.ConsultaOntologiaLinguistica \cap \forall Ejecuta_{Metodo}.GenerarArbolSemantico$
6.	$AnalisisProgmatico \sqsubseteq Tarea \cap \forall Ejecuta_{Metodo}.Interpretacion \cap \forall Ejecuta_{Metodo}.ConsultaOntologiaInterpretartiva$