VNIVERSIDAD NACIONAL AVTÓNOMA DE MEXICO

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

CONTEXTOS DEFINITORIOS EN LA EXTRACCIÓN DE TAXONOMÍAS: EL CASO DE PLN

T E S I S

QUE PARA OBTENER EL TÍTULO DE:
INGENIERA EN COMPUTACIÓN

PRESENTA:
ITA LUU YUYOCO CRUZ SHERLING

DIRECTOR DE TESIS:
DR. GERARDO SIERRA MARTÍNEZ

MÉXICO, D.F. 2013

Dedicatoria

A mi abuelita Esther Cruz Merino y a mi querido amigo Manuel Cabello Gutiérrez.

Agradecimientos

A mi madre, Patricia, y a mi padre, Refugio, por todo el esfuerzo y empeño que han puesto al cuidarme y educarme a lo largo de mi vida.

A mi familia por todos aquellos momentos de felicidad.

A mi hermano, Uriel por los momentos que nos reímos y distraímos de las ocupaciones cotidianas.

A mi novio Luis por su cariño y apollado durante todos estos años para seguir adelante.

A mi director, el Dr. Gerardo Sierra Martínez, por haberme dado la oportunidad de pertenecer al Grupo de Ingeniería Lingüística (GIL), por apoyarme y guiarme en el desarrollo de la tesis. Especialmente durante la recta final de este trabajo. También por transmitir el entusiasmo por la ingeniería lingüística.

Al Dr. Alfonso Medina por su apoyo y por incitarme a la investigación.

A la Dra. A María Potzi por sus consejos, apoyo y paciencia.

A mis sinodales por sus comentarios que me permitieron mejorar y fortalecer mi trabajo de tesis.

A Carlos Morales por su paciencia y comentarios de mi tesis.

A Laura confidente y amiga que conocí en el GIL. También Nicté, Brenda, Alejandro, Adrián, Octavio y al resto de GIL y a mis compañeros y amigos de la Facultad de Ingeniería.

Este trabajo se llevó a cabo con el apoyo del CONACyT, en el marco de los proyectos: Extracción de conocimiento lexicográfico a partir de textos de Internet con clave de registro 105711; El vocabulario básico científico de México: Una investigación de sus características, componentes y difusión con clave de registro 58923; Extracción de

relaciones léxicas para dominios restringidos a partir de contextos definitorios en español con clave de registro 82050.

A la máxima casa de estudios y mi segunda casa, la UNAM, por haberme dado una educación que vale oro; azul y oro. ¡México, Pumas, Universidad!

Por mi raza hablará el espíritu.

Resumen

En esta tesis se desarrolló e implementó una metodología basada en patrones léxicos para la extracción de taxonomías con corpus obtenido a partir de la web. El resultado obtenido fue una herramienta semi-automática flexible en la extracción de taxonomías en cualquier área del conocimiento.

Esta herramienta desarrollada con el método basado en patrones léxicos es capaz de extraer un corpus de forma automática a partir de un término dado por el usuario (llamado término semilla). También extrae todas aquellas frases que contengan el término semilla. Además realiza una extracción y recuperación de las frases que contienen el término semilla y presenten alguno de los patrones léxicos propuestos en esta tesis.

Para la evaluación del funcionamiento de la herramienta se programó una tarea para la clasificación de las frases obtenidas según el patrón léxico. Posteriormente, un experto realizó un análisis manual para validar los términos obtenidos de la clasificación en Procesamiento del Lenguaje Natural.

Por último, con los términos validados se realizó la graficación de la taxonomía para el caso específico del Procesamiento del Lenguaje Natural.

Índice general

Índice de figuras

Capítulo 1

Introducción

La Ingeniería en Computación se ha encargado en desarrollar sistemas y herramientas computacionales automatizadas y semi-automatizadas para facilitarles a los usuarios diversas tareas. Desde las convencionales como: el conteo de palabras, y en la realización de cálculos matemáticos; hasta las más complejas como: la navegación en internet, el mantenimiento de software, etc; y en la solución de nuevos problemas en la industria y academia.

Por la versatilidad en la aplicación de programas computacionales en diferentes áreas del conocimiento, la Ingeniería en Computación incursiona en áreas como la Lingüística con la finalidad de simplificar y mejorar sus procesos. Principalmente en el manejo de grandes cantidades de información digital y la comunicación humano-máquina. Auxiliándose de técnicas lingüísticas, estadísticas así como de métodos de Inteligencia Artificial y de Procesamiento del Lenguaje Natural (PLN).

Otra razón por la que el estudio de esta área ha incrementado su relevancia es debido al crecimiento exponencial de la información alrededor del mundo, actualmente gracias a internet, generándose una revolución digital. La cual da lugar a la era de la sociedad de la información, llamada así porque la información ha adquirido un alto grado de importancia en la actividad económica, al ser un nuevo modelo productivo [Salvat y Serrano, 2011].

Una muestra de la relevancia en la actividad económica es la economía basada en el conocimiento que emplea las Tecnologías de la Información y la Comunicación (TIC) mediante el almacenamiento y uso de la información con la finalidad de generar riquezas destinadas a mejorar la sociedad. [Salvat y Serrano, 2011, Correa, 2006].

Las TIC han propiciado el estudio de nuevas formas de accesar y generar conocimiento por medio de la inovación, las tecnologías y los recursos humanos para formar ambientes de negocios [Salvat y Serrano, 2011, Correa, 2006].

Por otra parte, el PLN procesa la información digital con métodos computacionales para diversas tareas como: la generación de ontologías; tesauros y taxonomías, supervisadas y no supervisadas. Para organizar y facilitar el acceso a la información, estas tareas se pueden implementar en la industria para la clasificación de productos y servicios, Gestión Financiera, Recursos Humanos, etc [INTECO, 2009].

Todas las formas en que los humanos organizan y representan la información se considera como el «mecanismo de generación conocimiento a partir del conocimiento almacenado.» [Pino, Gómez y De~Abajo, 2001, pg.~2].

Para organizar la información en la World Wide Web se han utilizado los Sistemas de Organización del Conocimiento (SOC), desde los tradicionales sistemas de clasificación y tesauros, hasta las más novedosas taxonomías, ontologías y redes semánticas [Díaz, Joyanes y Medina, 2009, Pino et al., 2001, Fernández, 2007].

Según Pino et al. (2001) los componentes del comportamiento inteligente son clasificados por la Enciclopedia de la Inteligencia Artificial[1], en las siguientes disciplinas como subáreas de la Inteligencia Artificial:

- Procesamiento de Lenguaje Natural
- Visión artificial
- Resolución de problemas
- Representación del conocimiento y razonamiento
- Aprendizaje
- Robótica

La clasificación y representación del conocimiento han sido unos de los problemas de estudio de la Estadística y la Inteligencia Artificial desde hace algunos años. Unos de los métodos que se han usado para estas tareas son las técnicas de «clustering» [2] y vocabularios controlados [3].

[1] Ver en la bibliografía como Encyclopedia of artificial intelligence, Volume 1

[2] Es una técnica en la minería de datos en la que objetos análogos se clasifican en un clúster (grupo). Tiene aplicaciones en máquinas de aprendizaje, procesamiento de señales e IA [Amini, Wah, Saybani y Yazdi, 2011].

[3] ver capítulo 2, Vocabularios Controlados.

1.1. Definición del problema

Debido a la gran cantidad de información contenida en la web que continuamente está en crecimiento, es necesario organizarla de cierta forma para que el acceso a ella sea más eficiente y así el análisis de sus datos por el usuario sea más fácil. Para ello es necesario describir y definir diversos términos que nos ayudarán a comprender mejor el funcionamiento de la representación del conocimiento como sistema de clasificación.

A partir de esta situación surge la necesidad de realizar una clasificación de los términos para las diferentes áreas del conocimiento con una metodología independiente del área del conocimiento al que se aplique. Por la razón anterior se realizarán programas computacionales basados en reglas lingüísticas que permitan obtener el mayor número de términos válidos en el menor tiempo, además de realizar su respectiva clasificación.

El método de representación elegido para esta tesis es la taxonomía porque se puede observar una jerarquización directa entre los términos padres e hijos. Este método es principalmente usado en los estudios academicos básicos para facilitar la organización de la información de acuerdo a su especialización. Además el PLN no cuenta con una taxonomía y esta investigación permitirá presentarla.

Pero el problema que persiste en nuestros días se encuentra en la construcción de taxonomías, en su mayoría de forma manual y aplicadas generalmente a las áreas biológicas. Cuando esta clasificación sigue una metodología independiente al área de conocimiento, puede generar taxonomías para cada una de ellas. Cabe señalar que el nivel de profundidad no será grande, sin embargo, la cantidad de términos a clasificar serán suficientes.

El interés por construir taxonomías automáticas para las tecnologías del lenguaje en los sistemas de recuperación y extracción de información tiene la finalidad de mejorar la eficiencia en éstos, al integrar taxonomías. Para el PLN algunos de sus intereses son obtener las terminologías automáticamente de un corpus y el estudio de la influencia de las relaciones léxicas o relaciones taxonómicas en los textos, que hacen que un término se encuentre relacionado con otro en un mismo dominio.

Las taxonomías facilitan el manejo y análisis de información, ya que brindan un orden en los datos, bajo una misma metodología y un mismo enfoque. Con técnicas de vocabularios controlados y lingüísticas se obtendrá una metodología para la extracción de taxonomías independientes del dominio del conocimiento.

En esta tesis se utilizarán las relaciones léxicas para extraer automáticamente la taxonomía del PLN en idioma español, mediante el método de extracción de relaciones léxicas, ya validadas en otras investigaciones como, en Ortega (2007).

1.2. Objetivos

- Diseñar una metodología supervisada aplicable a cualquier área del conocimiento para extraer taxonomías a partir de un corpus.
- Programar la metodología en un lenguaje de alto nivel para la extracción de taxonomías.
- Aplicar la metodología para el caso específico del PLN.
- Evaluar los términos obtenidos con un experto en el área.
- Presentar gráficamente la taxonomía extraída del PLN.

1.3. Estructura de la tesis

Esta tesis se encuentra formada por cinco capítulos. En el segundo capítulo se explicarán detalladamente los conceptos preliminares para comprender y diferenciar algunos vocabularios controlados, así como los elementos que los conforman. Además de describir la función de la extración y recuperación de información.

En el tercer capítulo se describirá la metodología para la extracción de taxonomías por etapas de elaboración. Primero se mostrarán los criterios para generar el corpus, después se explicará el método de extracción de información que se realizará y se mencionarán las formas en las que se puede presentar la taxonomía.

En el cuarto capítulo se mostrará el proceso realizado en este trabajo de investigación junto con las herramientas de programación utilizadas, el corpus generado y las pruebas efectuadas.

En el capítulo cinco se mostrarán los resultados obtenidos de la generación de la taxonomía del Procesamiento del Lenguaje Natural.

Por último, se mencionarán las conclusiones a las que se llegaron con esta tesis.

Capítulo 2

Bases sobre taxonomías

El humano desde hace siglos ha cuantificado y clasificado su entorno para modificarlo a su conveniencia. Una de las cosas que ha cuantificado y clasificado es el conocimiento; desde entonces el humano ha tratado de representar el conocimiento de diversas formas. Siendo la taxonomía una de las formas de representación más vieja que se ha estudiado junto con otras formas de representación del conocimiento tanto de forma lingüística como de forma informática.

Este capítulo se divide en dos secciones; en la primera sección se explican los conceptos básicos para contextualizar el problema y proporcionar la terminología necesaria que se usará en el siguiente capítulo. Es decir, se proporcinarán las definiciones preliminares con sus respectivas ejemplificaciones. Los conceptos que se explican en esta sección son: los aspectos sobre terminología; extracción y recuperación de información, centrándonos en extracción de relaciones léxicas; y en la segunda sección se explican los vocabularios controlados, hasta llegar al concepto de taxonomía, así como su definición, su estructura y los elementos que la conforman, donde también se muestran algunos ejemplos.

2.1. Terminología

Martí (2003) define a la terminología como el conjunto de términos usados en una disciplina. La terminología de una disciplina es un lenguaje especializado restringido a grupos de hablantes; por ejemplo, en Ingeniería en Computación se usan algunos términos como: compilador, grafos, computación paralela; en lengua y literaturas hispánicas: sintaxis, gramática; etc.

Hay que considerar que un término puede tener diferentes acepciones, es decir, un mismo término puede usarse en diferentes áreas del conocimiento para diferentes cosas, razón por la cual el término siempre será dependiente del contexto en el que sea usado.

Por otra parte, Cabré (1999) menciona tres acepciones para definir la terminología:

1. Según la disciplina, se ocupa de los términos especializados.
2. Según la práctica, son los principios que rigen la recopilación de términos.
3. Según el producto generado por esa práctica, es el conjunto de los términos de una disciplina especializada.

Nota. *Existen más formas de definir la términología, sin embargo, la definición que se emplea en esta tesis parte del enfoque científico-técnico, donde la terminología se concibe como «el conjunto de unidades de expresión y comunicación que permiten transferir el pensamiento especializado» según Cabré (1999: pg. 20). Considerando como objetivo el uso de términos para generar conocimiento.*

2.1.1. Términos y definiciones

Fedor (1995) define al **término** como la forma de nombrar a un concepto definido en un lenguaje especializado por una expresión lingüística, donde el término puede ser representado por una palabra, frase, icono, símbolo o acrónimo.

Según Fedor (1995), la **definición** es la repesentación de un término por declaraciones descriptivas y claras, donde se mecionan las características genéricas y específicas para hacer una diferenciación entre los términos relacionados.

Las definiciones se pueden dividir en tres categorías según Malaise et al. (2004):

- Definición no formal.
- Definición semi-formal.
- Definición formal.

De acuerdo con Trimble (1985), la **definición no formal** es la expresión en sentido general de un término, dicho de otro modo, es la descripción de las caracterícas comunes de un término que permite su entendimiento básico.

La **definición semi-formal** proporciona características específicas o atributos. [Meyer, 2001]

Y finalmente, la **definición formal** es la propuesta por Aristóteles consultada en Malaise et al. (2004) [1] que consta de un esquema:

$$X = Y + Z$$

- X es el término definido también llamado definiendum.
- = es la equivalencia relacionada.
- Y es la clase genérica a la cual pertenece el término (genus o género próximo).
- Z son las características específicas en las que X es diferente a otros elementos que componen la misma clase genérica (diferencia específica), es decir, rasgos propios que distinguen a dicho elemento con respecto a otros de su misma clase.

Ejemplo 1 (Definición aristotélica)**.** *Mesa: mueble, por lo común de madera, que se compone de una o de varias tablas lisas sostenidas por uno o varios pies, y que sirve para comer, escribir, jugar u otros usos [RAE, 2011].*

X = Mesa
Y = mueble
Z = por lo común de madera, que se compone de una o de varias tablas lisas sostenidas por uno o varios pies, y que sirve para comer, escribir, jugar u otros usos.

[1] Traducción literal realizada por mí

Partiendo de la definición formal Aristotélica, Sierra et al. (2006) consideran cinco tipos de definiciones, ver Fig.2.1:

- Definición analítica: presenta género próximo + diferencia. Por ejemplo.
- Exclusivamente género próximo: no proporciona una diferencia especfíca. Por ejemplo, «Java es un lenguaje de programación».
- Definición sinonímica: indica una equivalencia entre definiciones. Por ejemplo, «Medicina legal es también llamada medicina forense».
- Definición funcional: indica las funciones del concepto. Por ejemplo, «La computadora sirve para...».
- Definición extensional: menciona las partes que componen al concepto. Por ejemplo, «El sistema solar está conformado por los planetas Mercurio, Venus, Tierra, Marte, Júpiter, Saturno, Urano, Neptuno y Plutón».

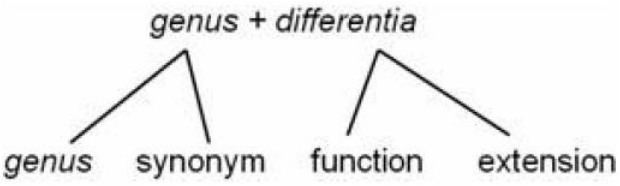

Figura 2.1: Tipología de las definiciones. Fuente: Sierra et al. (2006: pg. 3)

Nota. *Los términos no son elementos separados de las demás palabras, sino que coexisten de forma relacionada, por ello el análisis realizado en esta tesis toma en cuenta el uso de términos dentro de un texto.*

2.1.2. Relaciones léxicas

Los patrones lingüísticos han sido llamados de diferentes formas, pero la comunidad terminológica prefiere referirse a éstos «como conocimiento de patrones, con frecuencia se denominan patrones léxico-sintácticos». La relación léxica es la que se establece entre dicha palabra y otras, para definirla [Yule, 2007]. La representación de los patrones en algunos experimentos se restrigen a cadenas, especialmente, si la búsqueda es en Internet [Auger y Barriére, 2008].

Según Ortega (2007: pg. 14 y 15) las relaciones léxicas más comunes son:

- Sinonimia: se establece entre palabras diferentes cuando en un contexto tienen el mismo significado.
- Antonimia: se establece entre palabras cuyo significado es opuesto.
- Meronimia: se establece la relación entre las partes y el todo, generalmente presenta el siguiente patrón: *X es una parte de Y*.
- Hiponimia: es la relación entre una o más palabras que se encuentran incluidas semánticamente dentro de otra, es decir, el significado del concepto más específico (hipónimo) está incluido en el significado del concepto más general (hiperónimo).

Ejemplo 2 (WordNet)**.** *Es una gran base de datos léxica en inglés, donde los sustantivos, verbos, adjetivos y adverbios son agrupados en conjuntos de sinónimos cognitivos (synsets) que se encuentran vinculados por el significado del concepto semántico y las relaciones léxicas [Miller, Beckwith, Fellbaum, Gross y Miller, 1999, Miller, 1980], ver Fig.2.2.*

Figura 2.2: WordNet. Fuente: http://wordnetweb.princeton.edu/perl/webwn, consultada: 25/octubre/2012

Ortega et al. (2011) mencionan los patrones encontrados para la identificación de relaciones léxicas en un corpus con términos semillas [2]:

1. Para la relación de hiponimia
   - agua-líquido
   - gato-animal
   - manzana-fruta
2. Selección de semillas frecuentes
   - águila - ave
   - garza - ave
   - agamí - ave
3. Selección de semillas de diversos dominios
   - águila - ave
   - bulimia - enfermedad
   - arroz - cereal

En la siguiente lista se mecionan los patrones descritos en Ortega et al. (2011):

1. el <hipónimo>es el único <hiperónimo>
2. el uso de la <hipónimo>como <hiperónimo>
3. el <hipónimo>es uno de los <hiperónimo>más
4. de la <hipónimo>como <hiperónimo>de
5. de las <hipónimo>como <hiperónimo>
6. las <hipónimo>son una <hiperónimo>
7. el <hipónimo>es un <hiperónimo>que
8. el <hipónimo>es el <hiperónimo>que

[2]Los términos semillas son los términos que se tomaron como base para obtener las relaciones léxicas

9. de <hiperónimo>como <hipónimo>y
10. la <hipónimo>es un <hiperónimo>
11. la <hipónimo>una <hiperónimo>
12. las <hipónimo>son <hiperónimo>que
13. el <hipónimo>es un <hiperónimo>de
14. la <hipónimo>es la <hiperónimo>
15. la <hipónimo>es una <hiperónimo>que
16. la <hipónimo>como una <hiperónimo>
17. que la <hipónimo>es una <hiperónimo>
18. el <hipónimo>es una <hiperónimo>
19. la <hipónimo>es el <hiperónimo>de
20. de <hipónimo>y otras <hiperónimo>
21. del <hipónimo>como <hiperónimo>
22. el <hipónimo>es la <hiperónimo>
23. <hiperónimo>de <hipónimo>de
24. de <hipónimo>y <hiperónimo>
25. <hiperónimo>de <hipónimo>y
26. de <hipónimo>o <hiperónimo>
27. los <hipónimo>son <hiperónimo>
28. de <hipónimo>como <hiperónimo>de
29. el <hipónimo>y las <hiperónimo>
30. de los <hipónimo>y <hiperónimo>
31. de los <hipónimo>y los <hiperónimo>
32. la <hipónimo>es el único <hiperónimo>natural

33. <hiperónimo>de la actividad <hipónimo>y el deporte
34. la anorexia y la <hipónimo>son <hiperónimo>
35. de <hipónimo>y otros <hiperónimo>
36. el <hipónimo>es el <hiperónimo>de mayor longevidad
37. los <hipónimo>y otros <hiperónimo>
38. facultad de <hiperónimo>de la actividad <hipónimo>y
39. la <hipónimo>y otros <hiperónimo>
40. las <hipónimo>marinas son <hiperónimo>
41. el <hipónimo>es el <hiperónimo>interno más
42. licenciado en <hiperónimo>de la actividad <hipónimo>y del deporte
43. el <hipónimo>es el <hiperónimo>más grande del cuerpo

2.2. Extracción y recuperación de información

Desde hace algunos años la Extracción de Información ha sido estudiada por el área de Inteligencia Artificial (IA) para realizar extracciones de forma automática y así mejorar la eficacia de los resultados obtenidos en las consultas web.

Sierra (2009), Jackson y Moulinier(2007) y Grishman (2010) concuerdan en la definición de Extracción de Información (EI) como el proceso automático realizado por los sistemas computacionales aplicados a un conjunto de documentos electrónicos, también llamado corpus, escritos en lenguaje natural, para separar cierta información del resto de los documentos. Forma en que se analiza y descarta información irrelevante para posteriormente organizarla, automatizando así los procesos.

En otras palabras, el proceso de EI toma un conjunto de documentos electrónicos como entrada para ser analizados; el análisis es la consulta a los documentos de algún dato. Durante este análisis se obtienen datos específicos de los documentos y por último, los resultados son mostrados al usuario o almacenados en una base de datos [Sáez, 2009].

Para ello la EI primero determina el tipo de información que necesita extraer para cada uno de los dominios a tratar, a través de una plantilla o esquema definidos

para el problema. Es decir, el proceso de extracción consiste en analizar documentos, detectando para cada uno de ellos el o los tipos de plantillas a los que se ajustan y rellenar los campos de cada plantilla seleccionada con elementos extraídos del documento [Gonzalo y Verdejo, 2003].

Por lo tanto, los datos que pueden extraerse de un corpus son referentes a eventos, cifras o hechos particulares. Por ejemplo, un sistema de EI para encuestas de opinión de servicios puede obtener datos como el nivel de calidad, sugerencias del mismo y otros datos necesarios para la empresa. Es así como el análisis realizado por el sistema de EI puede extraer datos específicos.

Según Grishman (2010) las tareas que realiza la EI son:

1. **Extracción de nombres**: Identifica nombres en un texto y los clasifica como organizaciones, lugares, personas, etc.
2. **Extracción de entidades**: Identifica todas las frases referidas a un objeto de clases semánticas específicas y relaciones de frases referidas a otros objetos, mejor conocida como Reconocimiento de Entidades Nombradas.
3. **Extracción de relaciones**: Identifica pares de entidades con relaciones semánticas específicas. Una de las técnicas usadas por esta tarea es la extracción de contextos definitorios.
4. **Extracción de eventos**: Identifica instancias de eventos particulares y argumentos de cada evento.

Por otra parte, la **extracción de información terminológica y conceptual** es estudiada por la Ingeniería del Conocimiento (IC). Ya que el principal objetivo de esta área es la elaboración y organización de bases de conocimiento y la descripción de significados, mejor conocida como «conocimiento definitorio» [Sierra, 2009].

Del mismo modo la EI se ha enfocado en extraer el conocimiento definitorio a la par del contexto definitorio, siendo ésta la información que permite inferir el significado de los términos a partir de la descripción de sus atributos, características o relaciones semánticas [Sierra, 2009].

Recuperación de información

Al igual que la EI, la Recuperación de Información (RI) es el proceso informático para la adquisición, organización, almacenamiento y distribución de la información, según Jackson (2007). Es decir, es la búsqueda de información por medio de consultas a un corpus, base de datos, internet, etc.

Con respecto a lo anterior, la RI consiste en seleccionar automáticamente de una determinada colección de documentos, aquellos que se ajustan a una consulta realizada por el usuario. El resultado es una lista ordenada de los documentos (pueden ser textos, imágenes, voz, etc.) seleccionados de acuerdo con un criterio de relevancia [Gonzalo y Verdejo, 2003].

Ejemplo 3 (Sistemas de RI)**.** *Son los buscadores como Google, Yahoo, etc. que dan a los usuarios acceso a documentos a partir de una consulta realizada [Rueger, 2010], ver Fig.2.3.*

Figura 2.3: Sistema de recuperación de información

Según Martí (2003), se distinguen cuatro procesos básicos en la recuperación de información:

- Representación del texto (indiciación o indexación): Identificación de los términos que describen el contenido del texto.
- Representación de la consulta: Descripción y refinamiento de aquello que se busca en forma de consultas explícitas.
- Comparación de representaciones (recuperación de documentos): Se compara la consulta realizada en el conjunto de documentos para determinar el orden de relevancia en que estos se presentan.
- Evaluación de los documentos y realimentación de la búsqueda (relevance feedback): Se presentan los documentos al usuario y la información se utiliza sobre aquellos que verdaderamente le interesan para retroalimentar al proceso de búsqueda.

Nota. *El caso de los sistemas de recuperación automática de información se usan para indexar los documentos empleando índices o términos de los mismos [Gonzalo y Verdejo, 2003].*

Helen Brown, en 1957, manifestó en la Dorking Conference on Classification que El problema de la RI es transformar conceptos y sus relaciones, como se expresan en el lenguaje de los documentos, en un lenguaje más regularizado, con los sinónimos controlados y sus estructuras sintácticas simplificadas [Arano, 2005].

Medidas de desempeño en la RI

Una parte importante de la RI es su evaluación mediante algunas medidas de desempeño como: la precisión y la cobertura. Éstas evalúan los resultados obtenidos en los experimentos para determinar la robustez y la precisión de la predicción en la clase positiva (las variables mencionadas en esta sección se definen en la pg. 19).

La precisión es el número de casos recuperados correctamente que son relevantes para la búsqueda y se define como:

$$\textbf{Precisión} = \frac{tp}{tp+fp}$$

La cobertura es el número de casos relevantes para la consulta que se han recuperado correctamente y se define como:

$$\textbf{Cobertura} = \frac{tp}{tp+fn}$$

2.3. Extracción y recuperación de relaciones taxonómicas

En los últimos años, varias disciplinas científicas como la Lingüística Generativa, la Inteligencia Artificial, la Lingüística Computacional y la llamada Ciencia Cognitiva han mostrado un creciente interés en las múltiples facetas de las relaciones taxonómicas [Auger y Barriére, 2008].

Las relaciones taxonómicas son la base de cualquier sistema de representación del conocimiento y son claves para la generación de sistemas de procesamiento de información con capacidades semánticas y de razonamiento [Auger y Barriére, 2008].

Algunas de las relaciones taxonómicas o léxicas más estudiadas son: las de hiperonimia (o es-un) por Caraballo (1999), además de Ravichandran y Hovy (2002); meronimia (o parte-todo) analisada por Winston et al. (1987), Berland y Charniak (1999), Girju (2003), Pantel y Pennacchiotti (2006); las relaciones de definición estudiadas por Pasca (2005). Otras relaciones lingüísticas que se han estudiado es la de sinonimia por Lin et al. (2003), Baroni y Bisi (2004), así mismo por (Turney, 2001); y la de antonimia por Lucero et al. (2004) y Schwab et al. (2002).

Además, la relación de hiperonimia ha sido durante mucho tiempo el centro de interés para la generación de taxonomías y ontologías.

Según Ortega (2007) los métodos más usados para la extracción y recuperación de relaciones taxonómicas son:

- Métodos basados en diccionarios
- Métodos basados en agrupamientos
- Métodos basados en patrones

Métodos basados en diccionarios

Se centran en la extracción de información, como pueden ser las instancias de relaciones semánticas, a partir de diccionarios empleandolos como fuente de conocimiento estructurado. Este tipo de métodos son dependientes de la estructura del diccionario por lo que los hace muy precisos y facilita la obtención de la información. Principalmente estos métodos son enfocados a encontrar los hiperónimos de las palabras.

Ejemplo 4 (Definición del Diccionario de Inglés Contemporáneo Logman)**.** *primavera «La estación entre invierno y verano en la cual aparecen flores»*

En el ejemplo anterior, Ortega (2007) observa que *estación* es el hiperónimo de *primavera* por estar en la primera frase nominal de la definición, estableciendo en consecuencia una relación de hiponimia entre los términos *primavera* y *estación*. [Ortega, 2007]

Pero la gran desventaja de este método es que no presenta términos específicos de un dominio, debido a que la mayoría de las veces los diccionarios tratan términos generales del lenguaje. Este inconveniente propició el interés en la exploración de otros enfoques para la tarea de extracción de hipónimos [Ortega, 2007].

Amsler (1980) extrajo taxonomías de diccionarios en un formato legible computacionalmente (machine-readable) de dos diccionarios, Merriam-Webster Pocket Dictonary (MPD) y Seventh Collegiate Dictionary (W7).

Este trabajo parte del enfoque de que los diccionarios son fuente generadora de conocimiento taxonómico. Para extraer la taxonomía se basó en el índice de concordancias realizado por Jonh Olney en 1967 y desarrolló el siguiente procedimiento:

1. Elegir un dominio semántico a tratar, seleccionando el más representativo conjunto de verbos o sustantivos del dominio.

2. Usar la concordancia taxonómica para asegurar las definiciones mediante la asignación de los conjuntos elegidos como los núcleos de las definiciones.

3. Validar los significados de uso de las palabras y formar una lista de verbos o sustantivos.

4. Para cada elemento de la lista expandida, repetir los pasos 1-3 hasta que ninguno de sus elemento tenga algún uso en el diccionario y en la lista.

5. Generar una base de datos de los verbos desambiguados [3] y sus descendientes asociaciados de la lista.

6. Conectar de manera computacional la base de datos e imprimir su estructura como diagrama de árbol, detectando y enumerando todos los LOOPS que se presenten.

7. Resolver todos los loops presentados en el paso 6 uniendo los nodos del diagrama de árbol para formar sinónimos o reevaluar la fase de desambiguación del paso 5 para estos elementos.

[3] Los verbos desambiguados son el resultado de un proceso por el cual un verbo pierda sus distintas interpretaciones que no sean las que se plantean en el texto.

8. Mostrar la estructura final de los nodos conectados en el diagrama de árbol.

Dolan et al. (1993) trabajaron los diccionarios en línea para construir una base de conocimiento léxica estructurada. Mediante el uso del Diccionario de Inglés Contemporáneo Longman (LDOCE) para construir un grafo con el propósito de recuperar asociaciones semánticas entre palabras. El resultado entregado es una red interconectada de palabras unidas por arcos etiquetados con relaciones semánticas tales como *hiperónimo, parte de, ubicación y propósito.*

Métodos basados en agrupamientos

Estos métodos agrupan las palabras de acuerdo con su contexto, donde pueden asignarles una etiqueta a cada grupo [Ortega, 2007].

Pereira et al. (1993) construyeron una jerarquía de palabras no etiquetadas. La técnica consiste en evaluar experimentalmente un método de agrupación de las palabras según su distribución en determinados contextos sintácticos[4].

Riloff y Shepherd (1997) exponen un método basado en corpus que puede ser usado en la construcción de lexicones[5] semánticos para categorías específicas. Mediante el suministro de un conjunto de palabras semillas por cada categoría y un corpus de textos representativos, se obtiene una lista de palabras asociadas con cada categoría.

Cimiano et al. (2004) mostraron un método de clasificación conceptual basado en el análisis del concepto formal para la construcción de la taxonomía automática. Y una comparación entre los métodos: de la clasificación jerárquica aglomerativa, clasificación de división jerárquica (KMedias sección Bi como un algoritmo de división) y la comparación conceptual.

Métodos basados en patrones

Los métodos basados en patrones se sustentan de las convenciones o estilos que las personas repiten al momento de relacionar un hipónimo con su hiperónimo dentro de un texto, es decir, las convenciones pueden generalizarse en forma de patrones. Estos patrones permiten la extracción de instancias de la relación de hiponimia al aplicarse sobre un corpus [Ortega, 2007].

Malaise et al. (2004) usaron marcas léxicas y patrones para detectar al mismo tiempo definiciones y relaciones semánticas con 46 marcas y 74 patrones diseñados para un corpus de antropología. Reportando una evaluación en un segundo corpus en el área de la dietética del 4 % al 36 % de cobertura y del 61 % al 66 % de precisión.

[4]Se refieren al significado de los términos de acuerdo al contexto en el que se maneja.
[5]Los lexicones son el conjunto de palabras de un lenguaje.

Según Auger y Barriére (2008), el modelo basado en la extracción de relaciones semánticas con frecuencia implica cuatro pasos principales:

1. Definición de la relación semántica de interés.
2. Descubrimiento de patrones reales que expresen explícitamente en el texto dicha relación, así como las condiciones sintácticas en las que el significado de la relación específica es realizada.
3. Búsqueda de instancias relacionadas con los patrones.
4. Estructuración de las nuevas instancias como parte de una nueva ontología[6], o una existente (o base de datos terminológica).

Medidas de desempeño del extractor taxonómico

A partir del enfoque de la matriz de confusión 2.1, se evaluarán los resultados obtenidos en los experimentos de esta tesis.

	Clasificado como positivo	**Clasificado como negativo**	**Total**
Realmente positivo	tp	fn	p
Realmente negativo	fp	tn	n
Total	p'	n'	N

Tabla 2.1: Matriz de confusión

p –>Es el número de casos positivos reales.

n –>Es el número de casos negativos reales.

p' –>Es el número de casos clasificados como positivos

n' –>Es el número de casos clasificados como negativos.

tp y **tn** –>Son las clasificaciones correctas.

fn y **tp** –>Representan los dos tipos de errores.

[6]Ver página 28.

«Para un caso realmente positivo, si la predicción es también positiva, se le denomina positivo verdadero, tp (true positive). Si la predicción es negativa, para el mismo caso positivo, se le llama falso negativo, fn (false negative). Si el caso es realmente negativo y se predice como positivo, es un falso positivo, fp (false positive). Si con el mismo caso negativo, se predice negativo, se obtiene un negativo verdadero, tn (true negative).» [Soriano, 2011]

2.3.1. Extracción de contextos definitorios

Un contexto definitorio es el fragmento textual extraído de un documento de una disciplina específica el cual aporta información para conocer el significado de un término dentro de la discplina a la que pertenece el documento.[Alarcón, Bach y Sierra, 2007]

Según Sierra (2009), del contexto definitorio se estudian dos áreas:

- **La extracción de relaciones semánticas**: Obtiene las relaciones de hiperonimia, hiponimia, holonimia, meronimia, sinonimia, etc. entre las palabras.
- **La extracción de contextos definitorios**: Obtiene las frases de un documento que cumple con las características de un contexto definitorio.

En la fig. 2.4 el contexto definitorio (CD) según [Alarcón et al., 2007] se encuentra formado por:

- T: Término
- D: Definición
- PD: Patrón definitorio
- PP: Patrón pragmático

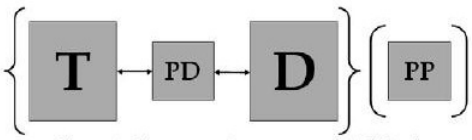

Figura 2.4: Estructura de un contexto definitorio. Fuente: Sierra (2009: pg.17)

En la figura anterior se muestra la relación existente entre el término y la definición a través del **patrón definitorio**, fundamental para encontrar el contexto definitorio. Sin embargo, el patrón prágmatico puede aparecer o no, porque es una estructura que aporta información extra al término. Por ejemplo, *en términos generales* o *en esta investigación.*[Alarcón et al., 2007]

Ejemplo 5 (Contexto definitorio etiquetado según sus partes)**.** *<PP>Tradicionalmente </PP>, <T>la logística </T><PD>se define como </PD><D>el arte militar que estudia el movimiento, transporte y estacionamiento de las tropas fuera del campo de batalla</D>.[Sierra, 2009]*

En el ejemplo anterior el autor conecta el término con la definición mediante el uso de *se define como* y el uso del patrón pragmático *tradicionalmente*, que en este caso indica un matiz especial sobre el significado del término.[Sierra, 2009]

Debido a que el patrón pragmático no siempre puede existir en los CDs, no se trabajará con esta sección del contexto definitorio en esta tesis.

Por lo tanto, el patrón definitorio, además de unir el término con la definición muestra una relación léxica que permite encontrar los hipónimos o hiperónimos del mismo término dentro de una disciplina. Es decir, por medio de los patrones definitorios se puede encontrar la terminología de una disciplina y su jerarquía.

Ejemplo 6 (Uso de las relaciones léxicas)**.** *«En la actualidad, la Computación Cuántica Geométrica (CCG) es un área activa de investigación el contexto de la Computación Cuántica»...[Sicard y Vélez, 2005, pg.~6]*

El ejemplo seis es un fragmento de un CD el cual muestra el patrón definitorio *es un* revelando la existencia de una relación de hiponimia o hiperonimia entre el término y su definición, donde el hipónimo es *Computación Cuántica Geométrica (CCG)* y el hiperónimo es *área*. En ambos términos se indica la jerarquía entre ellos por medio del patrón definitorio.

Un elemento clave en el proceso para reconocer CDs de forma automática es la identificación de los patrones léxicos que se emplean para relacionar al término con su definición [Sierra, 2009].

Sierra (2009) menciona dos tipos generales de patrones definitorios fig.2.5 para el español: tipográficos y sintácticos.

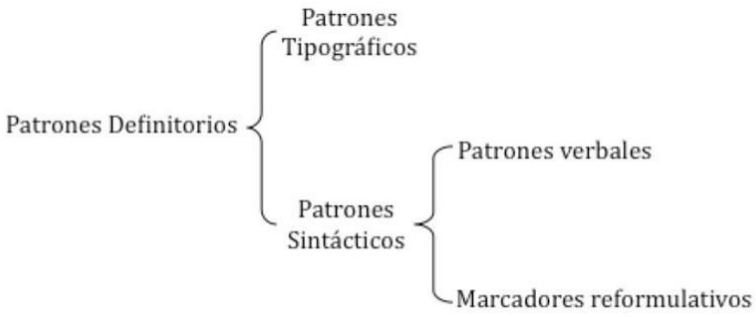

Figura 2.5: Tipología de patrones definitorios. Fuente: Alarcón (2009: pg.127)

Patrones tipográficos: son aquellos elementos utilizados para resaltar un término en un texto de forma visual.

Ejemplo 7 (patrones tipográficos).

- *«Un **Operador** es un símbolo o nomenclatura al cual se le da un significado predeterminado.»[Osorio, 2008, pg.~47]*

- *«**Identidad**: Es la propiedad característica que tiene un objeto que le distingue de todos los demás»[Osorio, 2008, pg.~39]*

- *«A la organización de los datos en el programa se le concoce como* estructura de datos*»[Osorio, 2007, pg.~13]*

El ejemplo siete muestra tres tipos de patrones tipográficos, el primer de estos ejemplos muestra el uso de letras en negrillas para resaltar el término «Operador» del resto del texto. En el segundo ejemplo el autor resalta el término «Identidad» en negrillas seguido de dos puntos y en el último ejemplo se resalta el término «estructura de datos» con otro tipo de letra al resto del texto. Éstos son algunos ejemplos de los formatos de patrones tipográficos que existen.

Patrones sintácticos: son verbos que conectan el término con la definición y estos se dividen en patrones verbales y marcadores de reformulación.

Patrones verbales definitorios (PVD) o **patrones verbales** son los conectores que tienen como núcleo un verbo. Algunos de estos verbos son comúnmente considerados como verbos metalingüísticos [Sierra, 2009], porque siendo los verbos pertenecientes al lenguaje, definen al mismo lenguaje. Ejemplo: *definir, entender o denominar.*

Después de analizar el lenguaje general en diferentes situaciones también se encontraron verbos los cuales no expresan siempre definiciones, como los verbos *ser* y *considerar* por separado. Para ambos tipos de verbos ocurren dos tipos de construcciones sintácticas verbales [Sierra, 2009]:

1. Verbo de forma aislada: entendemos o definimos
2. Partícula gramatical: se

Marcadores reformulativos: se utilizan para explicar el significado de un término que permite retomar un elemento de un discurso para explicarlo de otra manera ayudando a la cohesión textual.

Algunos ejemplos de marcadores reformulativos son: por ejemplo, es decir, esto es, en otras palabras, dicho de otra manera.

Identificación de límites de los CDs

Por último, lo que falta explicar son los límites de los CDs, es decir, dónde comienza y dónde termina un CD en un texto, ya que en la mayoría de los casos dentro de un documento se presentan de diferentes formas. Razón por la cual es necesario identificar de forma precisa los términos relacionados con los patrones definitorios, además de encontrar su respectiva jerarquía.

Una de las formas básicas para saber dónde inicia y termina un CD se determina de punto a punto, pero se ha observado que los resultados no siempre son buenos porque no siempre la definición termina después de un punto. [Hernández, 2009]

Ejemplo 8 (Contexto definitorio)**.** *«La pulgada (qué se designa como inch o in, por su nombre inglés) se define precisamente como 2.54 centímetros (cm; 1 cm=0.01m).» [Giancoli, 2007, pg.~385]*

Ejemplo 9 (Contexto definitorio II)**.** *«La física es una ciencia que se encarga de estudiar e interpretar los fenómenos que ocurren en la naturaleza. Esta ciencia*

se expresa a través de una variedad de lenguajes, de los cuales históricamente el matemático constituye uno de los más utilizados por la comunidad de los físicos.» [Miranda, Lobo, Castro, Mendoza y Gracerant, 2010, pg.~1]

En el ejemplo ocho podemos ver que la definición del término **pulgada** no acaba en el punto decimal. En el ejemplo nueve también podemos observar que la definición del término **física** no termina en el punto y seguido.

Con la finalidad de evitar extraer información que no sea parte del CD y mejorar el sistema de extracción, es necesario plantear reglas lingüísticas que permitan delimitar las definiciones automáticamente cuando éstas terminan antes o después del primer punto.

Para generar la taxonomía de forma semi-automática en el estudio de CDs, nos enfocaremos en el área de extracción de relaciones léxicas. Usando principalmente patrones léxicos (patrones verbales definitorios y marcadores reformulativos) más adelante se explicará a detalle este elemento.

2.4. Vocabularios controlados

Los *lenguajes controlados* también conocidos como *vocabularios controlados* (VCs) son un conjunto de palabras o frases de un tema específico (términos de un área), estandarizados estos elementos por alguna comunidad científica. Donde tratan de organizar la información y eliminar las ambigüedades dentro de las mismas definiciones y así encontrar una interpretación precisa de los elementos usados [Harpring, 2010], [Aude y Soto, 2006]. Por ejemplo, los VCs más comunes son los diccionarios y glosarios.

Schwitter (1999) define a los VCs como subconjuntos de los lenguajes naturales cuyas gramáticas y diccionarios se han restringido con el fin de reducir o eliminar la ambigüedad y la complejidad. Tradicionalmente, los lenguajes naturales controlados se dividen en dos grande categorías: aquellos que mejoran la facilidad de lectura para los humanos, en particular para los los hablantes no nativos, y los que mejoran el procesamiento computacional de un texto.

Posteriormente, Schwitter (2010) define el término vocabulario controlado como la brecha que existe entre el lenguaje natural y el lenguaje formal mediante el uso de lenguajes naturales controlados.

Deacuerdo con la norma ANSI/NISO (National Information Standars Organization de EE.UU.) Z39.19-2005 los vocabularios controlados se utiliza para mejorar la efica-

cia del almacenamiento y recuperación de la información en sistemas de navegación web y otros ambientes de búsqueda para identificar y localizar el contenido deseado a través de algún tipo de descripción mediante el lenguaje. El propósito principal de los vocabularios controlados es lograr la coherencia en la descripción de los objetos de contenido y facilitar la recuperación[7].

Según Harpring (2010), el propósito de los VCs es organizar, recuperar la información y proporcionar a través de la terminología, siendo de mucha utilidad para mejorar las consultas de información en sistemas web.

Lancaster (2002) y Harpring (2010) explican que las principales funciones desempeñadas por los VCs son:

- Reducir ambigüedades semánticas mediante la reunión de los sinónimos de los términos.
- Mejorar la representación del tema específico a través de vínculos y relaciones recíprocas para asegurar la catalogación y recuperación de la información.
- Facilitar la realización de búsquedas amplias.

Según Centelles (2005) los principales tipos de VCs son: las listas, los anillos de sinónimos, las taxonomías y los tesauros.

De acuerdo con la norma ANSI/NISO Z39.19-2005 definen a las listas y anillos de sinónimos como:

- Lista: Un conjunto de palabras o frases que se muestran en una forma organizada.
- Anillo de sinónimos: Un conjunto de palabras o frases que se consideran equivalentes para los fines de recuperación de información.

Sin embargo, éstos no son los únicos VCs. Existen otros como folksonomías y ontologías (se explican en las siguientes páginas). Además los VCs se pueden ordenar de acuerdo a su nivel de complejidad como se puede observar en la fig. 2.6.

[7]Traducido por mi.

Figura 2.6: Complejidad de las herramienta para la representación de conocimiento. Fuente: Soler y Leiva (2010) a partir de la Norma Z39.19:2005, pg.17

2.4.1. Tesauros

El término Tesauro proviene del latín thesaurus y del griego thesaurós, que significa tesoro o repositorio de palabras [Arano, 2005].

Según Centelles (2005), el tesauro es un conjunto de palabras o frases con términos equivalentes explícitamente identificados como se puede observar en la fig. 2.7 y con palabras o frases ambiguas (por ejemplo: homógrafos) que los hacen únicos. Este conjunto de términos también puede incluir otras relaciones más generales o restringidas.

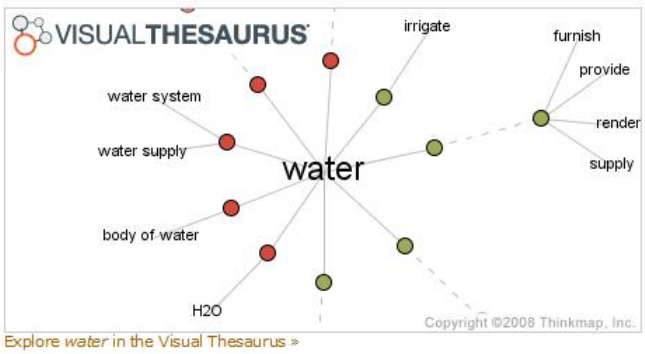

Figura 2.7: Tesauro. Fuente: http://thesaurus.com/, consultada: 25/octubre/2012

Aún cuando el significado del término Tesauro ha evolucionado a lo largo del tiempo, éste ha conservado algunas caracaterísticas generales. Coincidiendo Soler (2009), Miranda (1995) y Arano (2005) en definirlo como una lista alfabética de términos de una disciplina, relacionados entre sí por medio de reglas heurísticas o intuitivas.

Gimeno (2004) define funcionalmente al tesauro como «instrumento de control terminológico utilizado al trasladar a un lenguaje más estricto la lengua natural empleada en los documentos y por los indizadores o los usuarios».

Deacuerdo con la norma ANSI/NISO Z39.19-2005 el tesauro se define como un VC organizado de tal forma que las relaciones entre los térmios se identifiquen claramente por medio de los indicadores de relación.

Miranda (1995) señala tres etapas para la elaboración de tesauros:

1. «Recopilación de las terminologías sobre la materia.»
2. «Reducción de éstas al mínimo.»
3. «Establecimiento de una gramática que relacione los términos seleccionados.»

Estructura del tesauro según Miranda (1995):

- Descriptor
- Notas de alcance
- Términos relacionados
- Referencias de *véase*

Nota. *El descriptor es el núcleo del tesauro cuyo término está compuesto por una o más palabras que hacen referencia a un concepto, es decir, son los términos o conceptos de una disciplina. Se diferencia de las palabras-clave o de los unitérminos en que estos forman parte del lenguaje natural, de la indización o recuperación en lenguaje libre, mientras que el descriptor tiene que estar admitido como término de indización en el tesauro.[Gimeno, 2004]*

2.4.2. Ontologías

Es una rama de la Filosofía que tiene por objetivo explicar de forma sistemática la estructura de los objetos (relaciones, propiedades y procesos). Entre las décadas de los 80's y 90's el término ontología incursiona en el área de la IA como forma de representación del conocimiento. A partir de ello se han realizado estudios en este campo del conocimiento[Soler, 2009].

El término ontología actualmente se ha desarrollado en los ámbitos de la ingeniería del conocimiento, el PLN, los sistemas coperativos de información, integración inteligente de información y gestión del conocimiento como software para la administración de información [Smith, 2001].

Quero (2007), Aufaure y Moulon (2006) concuerdan en que la ontología es una forma estructurada de explicar el conocimiento de un dominio, es decir, usa un vocabulario perteneciente a un dominio, el cual consiste en la descripción jerárquica de cada uno de los conceptos. El vocabulario se define en: entidades, clases, propiedades, predicados, funciones y la relación entre estos componentes.

En otras palabras, «las ontologías se encargan de definir los términos utilizados para describir y representar un área de conocimiento.»[Díaz et al., 2009]

Además, Díaz et al. (2009) afirman que las ontologías son «herramientas que sirven para estructurar conceptualmente determinados ámbitos del conocimiento por medio de vocabularios controlados, proporcionando una descripción lógica y formal que puede ser interpretada tanto por las personas como por las máquinas.»

Peis et al. (2003) mencionan que una ontología «estará formada por una taxonomía relacional de conceptos y por un conjunto de axiomas o reglas de inferencia mediante los cuales se podrá inferir nuevo conocimiento».

«Las ontologías son construcciones formales que representan nodos conceptuales y expresan las relaciones conceptuales que establecen entre sí. Su complejidad y su atomización suele ser mayor que en el caso de los tesauros, ya que su finalidad no es clasificar documentos y localizarlos, sino que se construyen con el fin de ordenar y relacionarlos con las expresiones lingüísticas que los vinculan».[Lorente, 2005]

Una traducción literal de Aufaure y Soto (2006) menciona que las ontologías consisten generalmente en una taxonomía o vocabulario y reglas de inferencia tales como transitividad y simetría.

Las reglas de inferencia pueden ser usados en conjunción con RDF[8] o Topics Maps[9], por ejemplo, para permitir la validación de la consistencia o para inferir información nueva.

Según Castillo et al. (2010), los componentes de las ontologías son:

- Atributos: «Los atributos representan la estructura interna de los conceptos. Según su origen, los atributos se clasifican en: específicos y heredados». [Castillo, Franco y Giraldo, 2010, pg.~369]
- Conceptos: Son aquellas que ideas que se formalizan. Estos pueden ser clases de métodos, estrategias, objetos, entre muchos otros. En las ontologías los conceptos son usados como base para describir el conocimiento.
- Relaciones: Son la forma en que se muestran los enlaces entre los conceptos del dominio, permitiendo la interacción entre ellos y la posibilidad de formar la taxonomía (para información de taxonomías en). Por ejemplo: *subclase-de, parte-de, parte-exhaustiva-de, conectado-a*, etc.
- Funciones: Son un tipo de relación que identifica un elemento particular entre varios elemento de la ontología. Por ejemplo, pueden aparecer funciones como *jerarquización-clase, asignar-fecha*, etc.
- Instancias: Se utilizan para representar objetos determinados de un concepto como pueden ser los miembros de una clase. Las instancias no pueden ser divididas sin perder su estructura y características funcionales. Por ejemplo, la instancia «perro».
- Axiomas: Sentencias que siempre se aceptan como verdaderas. Los axiomas son empleados para validar la ontología.

2.4.3. Folksonomías

De acuerdo con Moreiro (2006) fue Thomas Vander Wal quien acuñó el término folksonomía referido a una *clasificación gestionada popularmente*, al combinar los términos *folk* (gente, popular) y taxonomía (gestión *taxis* de la clasificación *nomos*).

[8]Resource Description Framework, Marco de Descripción de recursos es un modelo estándar para el intercambio de datos en la Web.

[9]Los Topic Maps son un estándar para la web que posibilita la navegación conceptual [Moreiro, Sánchez y Morato, 2003].

La folksonomía o folcsonomía se define como el sistema de clasificación colaborativo del contenido web, generado por los internautas. Se genera mediante palabras clave usadas como etiquetas en un sistema de etiquetado o clasificacion de objetos web no jerárquico, cuya gestión se realiza por un sistema automático [Díaz et al., 2009, Moreiro, 2006, Noruzi, 2007].

Ejemplo 10 (Folksonomía)**.** *Wikipedia es una enciclopedia gratuita, multilenguaje, online, construida con la colaboración de voluntarios. En esta enciclopedia las categorías se organizan en una estructura similar a la taxonomía, donde cada categoría puede tener un número arbitrario de subcategorías establecidas a causa de hiponimia o meronimia.[Zesch y Gurevych, 2007]*

Según Noruzi (2007) los sistemas más populares basados en folksonomías son:

- Del.icio.us: www.del.icio.us
- CiteULike: www.citeulike.org
- Connotea: www.connotea.org
- Flickr: www.flickr.com
- Furl: www.furl.net
- LibraryThing: www.librarything.com
- Scuttle: www.scuttle.org
- Shadows: www.shadows.com
- Simpy: www.simpy.com
- TagCloud: www.tagcloud.com
- Tagzania: www.tagzania.com
- Technorati: www.technorati.com
- Unalog: www.unalog.com
- Yahoo's MyWeb: http://myweb.yahoo.com
- YouTube: www.youtube.com

2.4.4. Taxonomías

Etimológicamente taxonomía procede de los términos griegos «taxis», ordenación, y «nomos», norma.

«Aristóteles fue uno de los primeros en utilizar este término, hacia el año 300 A.C., para designar esquemas jerárquicos orientados a la clasificación de objetos científicos.»[Díaz et al., 2009, pg.~245]

Ejemplo 11 (Taxonomía)**.** *En la fig.2.8 se puede observar el Árbol de Porfirio, el cual es la representación gráfica más famosa de las categorías de Aristóteles, que consiste en la forma de un árbol de relaciones [Sowa, 2000]. Este árbol muestra la «relación de subordinación de la substancia considerada como género supremo a los géneros y especies inferiores hasta llegar al individuo ».[Ferrater, 1969]*

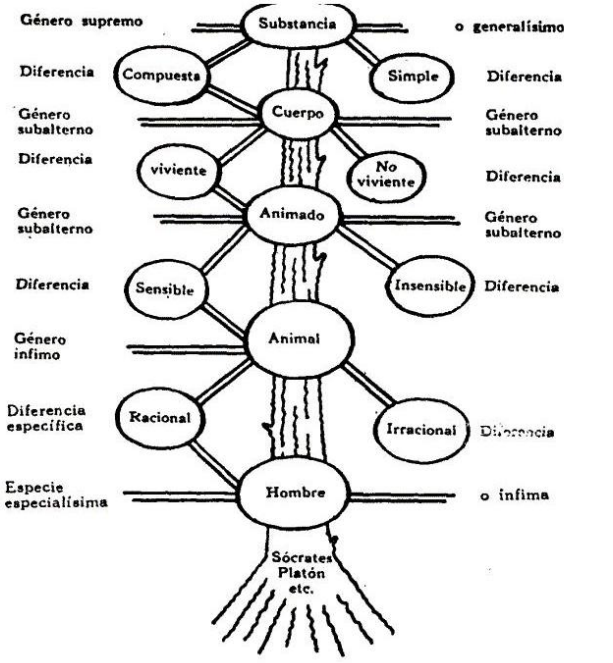

Figura 2.8: Árbol de Porfirio. Fuente: Ferrater (1969: pg. 126)

Posteriormente, en el siglo XVIII, Carlos Linneo, también conocido como Carl von Linné o Carolas Linnaesus, se convierte en el Padre de la Taxonomía por haber inventado el sistema de clasificación binario de los seres vivos, el cual sigue vigente. [Ramírez, 2007]

«Hoy en día los principales usuarios de las taxonomías son las grandes agencias internacionales de inteligencia (CIA, FBI, M16, etc.), que utilizan desde hace muchos años tecnologías de rastreo de información, clasificación y recuperación muy sofisticadas.»[Díaz et al., 2009, pg.~247]

Según Centelles (2005) «Una taxonomía es un tipo de vocabulario controlado en que todos los términos están conectados mediante algún modelo estructural (jerárquico, arbóreo, facetado...) y especialmente orientado a los sistemas de navegación, organización y búsqueda de contenidos de los sitios web.»

Según Ramírez (2007) define la taxonomía como la disciplina científica que tiene la labor de clasificar y jerarquizar sistemáticamente los organismos de acuerdo con los rasgos o carácteres que comparten, entendiéndose como clasificar el reconocer, nominar y agrupar.

El propósito de la taxonomía es desarrollar un ordenamiento lógico de los elementos basándose en su afinidad.[Montoya, 2008]

Requisistos generales que deben cumplir los atributos de los objetos a la hora de categorizarlos en la taxonomía según Díaz et al. (2009) son:

1. Objetivo: Identificación del atributo en un dominio del conocimiento.

2. Determinación: Proceso claro para indentificar el atributo.

3. Repetición: Coincidencia entre diferentes personas en el valor de los atributos de un objeto observado.

4. Mutuamente excluyente: Un grupo determinado de atributos solo debe pertenecer a una categoría.

5. Exhaustividad: Cuando los grupos incluyen todas las posibilidades.

6. Aceptabilidad: El atributo categorizado debe ser lógico y validado por especialistas en el dominio del conocimiento .

7. Utilidad: El atributo puede ser usado para generar conocimiento.

Del término taxonomía se desprende el término *taxonomía corporativa*, el cual es la organización (indizar o clasificar el contenido web) de diferentes tipos de contenidos digitales como aplicaciones informáticas, datos no estructurados, etc. en diversas organizaciones para generar capital. Esta forma de organizar la información facilita la navegación por buscadores, sistemas de filtraje, aplicaciones de minería de datos, etc. [Argudo y Centelles, 2005]

Al aplicar una taxonomía al contenido de los sitios web se estructuran para facilitar a los usuarios la navegación y la búsqueda de información. Por lo tanto, las taxonomías son una parte fundamental para jerarquizar y organizar los contenidos que se presentan en un sitio [Fernández, 2007].

Nota. *La forma de representación del conocimiento que se eligió es la taxonomía porque que se requería clasificar los términos asociados al Procesamiento del Lenguaje Natural de acuerdo a una jerarquía de uso de éstos en el dominio del conocimiento al que pertenecen.*

Centelles (2005) explica el proceso para la construcción de las taxonomías corporativas:

1. Acotar la realidad que será representada por la taxonomía. La delimitación puede ser por área del conocimiento, entidad, etc.
2. Extraer del conjunto de términos o categorías que representan dicha realidad.
3. Control de términos o categorías. Es la validación de los términos.
4. Establecimiento del esquema y la estructura de la organización de los términos o categorías: Se determinan los criterios para dividir o agrupar categorías (temas, materias o disciplinas, etc). Dentro de los criterios de agrupamiento puede utilizarse el modelo estructural, que define la relación que se establece entre los grupos de categorías producidos del esquema de organización. Otro modelo aplicado es el jerárquico (basado en la relación «tipo de») y el arbóreo (basado en la relación parte de).

Nota. *Se empleará el proceso de construcción de taxonomías corporativas para la construcción de la taxonomía del PLN. Debido a que la metodología se ajustaba a los requerimientos de la investigación.*

Los componentes básico de una taxonomías son :

- Taxones: Son los elementos que conforman la taxonomía a cualquier nivel taxonómico. Los taxones o taxas también son conocidos como *categorías taxonómicas*. [Montiel, 1991] y [Montoya, 2008]
- Relaciones léxicas (cf. supra, p. 8)

Nota. *En nuestro caso, el taxón o término semilla es el término perteneciente al área de investigación a clasificar y las relaciones léxicas son los patrones que muestran una relación de género próximo, sinonímia o diferencia específica entre los términos.*

2.4.5. Diferencias entre los vocabularios controlados

En las tablas 2.2 y 2.3 [10] se muestra una comparación entre los VCs (Folksonomías, Taxonomías, Tesauros y Ontologías) mencionados en este capítulo.

[10] Los operadores matemáticos que aparecen en la tabla muestran una comparación de los niveles de complejidad que presentan los diferentes VCs.

	FOLKSONOMÍAS	**TAXONOMÍAS**	**TESAUROS**	**ONTOLOGÍAS**
Objetivo	Organización del conocimiento. Categorizar y agrupar información	Organización del conocimiento. Clasificar información	Organización del conocimiento. Indizar y recuperar información	Organización del conocimiento. Sistematizar y explotar el conocimiento
Origen	Década de 2000 (Entorno digital)	Reusadas en la década de 1990 (Entorno digital)	Década de 1950 (Entorno analógico y digital)	Década de 1980 (Entorno digital)
Productores	Cualquier usuario de la web	Personal especializado en la materia en donde se aplican	Profesionales de la información y la documentación	Profesionales de la TIC
Lenguaje	Lenguaje propio del internauta (natural y dinámico)	Terminología comprensible por los usuarios	Terminología consensuada, normalizada y controlada	Lenguaje natural, lenguaje controlado y lenguaje formal
Estructura	No	Jerárquica	Sistemática o macrotesauro, jerárquica, alfabética, índice (Kwic ó Kwoc)	Taxonomía, tablas con conceptos, sinónimos, descripciones, instancias, relaciones, atributos, valores, axiomas
Coste de elaboración	-	+	++	+++

Tabla 2.2: Diferencias entre los VCs; Fuente: Concha Soler Monreal et al. (2010: pg.377)

Actualización	Inmediata y arbitraria por los propios internautas	Inmediata y arbitraria por los diseñadores de los contenidos webs	Periódica y consensuada según la evolución de la terminología	Inmediata según la evolución del contenido de la propia ontología
Tipos de relaciones	Asociativas	Jerárquicas	Jerárquicas, asociativas y de equivalencia	Jerárquicas, asociativas, de equivalencia y cualquier otro tipo (temporales, familiares, causas-efectos, síntomas-tratamientos, etc.)
Aplicaciones	Navegación y búsqueda en la web social	Presentación y organización de contenidos digitales	Indización, navegación y búsqueda	Organización, clasificación, navegación, búsqueda, interoperabilidad y razonamiento automático
Normativas	No hay estándar oficial	No hay estándar oficial	ISO 25964	No hay estándar oficial
Presentación	En forma de conjuntos de etiquetas o tags	Jerárquica	Alfabética, jerárquica, sistemática	Recomendaciones de W3C
Inclusión de definiciones	No	No	Sí	Sí
Propiedades de los términos	No	No	No	Sí
Control de la ambigüedad	No	+	++	+++
Control de la sinonimia	No	+	++	+++
Editores para su construcción	No	No	MultiTes, Stride TCS, Lexico, TermTree 2000	Protegé, Ontolingua, OntoEdit
Coste de mantenimiento y actualización	-	+	++	+++

Tabla 2.3: Diferencias entre los VCs; Fuente: Concha Soler Monreal et al. (2010: pg.377)

Capítulo 3

Metodología para la extracción de taxonomías

En este capítulo se realiza una descripción general del método propuesto para extraer semi-automáticamente la taxonomía asociada al PLN, es decir, se usará un método supervisado. Debe recordarse que una taxonomía se encuentra formada por relaciones léxicas y taxones.

El método que se propone en este trabajo consiste en la recuperación basada en patrones de parejas de relaciones léxicas de hipónimos-hiperónimos y sinónimos, a partir de textos extraídos de la web.

Algunos de los patrones usados para extraer las relaciones léxicas para este trabajo se tomaron de Ortega (2007) y otros son propuestos a partir de un análisis de los textos obtenidos. El uso de relaciones léxicas se eligió para la construcción de la taxonomía debido a que Ortega (2007) menciona que «tiene una alta capacidad de extraer correctamente un par de palabras que mantengan la relación deseada».

Se tomará el proceso propuesto por Centelles (2005) para la construcción de taxonomías corporativas, ya que este es un proceso general que pueden seguir todas las construcciones de taxonomías.

Primero, la limitación de la realidad se centrará en el dominio del conocimiento sobre PLN, para ello se usará tal cual este término en la extracción y recuperación de información. Porque su uso en los textos engloba sus características, las formas como es conocido, etc; entonces, cuando se presente el término a buscar (término semilla), se encontrarán los términos relacionados.

A continuación, para extraer la taxonomía del PLN se compilará un corpus relacionado con este término. Es decir, se realizará una recuperación de la información a través del buscador de Google y su extracción de forma automática mediante un programa en Python.

Después se realizará un análisis al corpus generado para encontrar las parejas de patrones léxicos que contengan el término de PLN; de esta forma se aplicará un control terminológico. El análisis que se usará es automático para la extracción de frases que contengan los patrones propuestos mediante un programa en Python, desarrollado para esta tarea. Acto seguido, un experto en el dominio del conocimiento del término evaluará las frases extraídas y decidirá si el término extraído es un término o no.

Por último, se establecerá el esquema de representación para los términos encontrados, según su jerarquía, a través del uso del software Graphviz [1]. Recuerde, las únicas relaciones que se usarán son las de hiponímia, hiperonímia y sinonímia.

La taxonomía resultante consistirá en tres categorías taxonómicas, es decir, estará conformada por tres niveles de profundidad: el superior, el inferior y el del término semilla. Esto se debe a que el programa realizado en Python sólo extrae información de las relaciones léxicas de hiponímia-hiperonímía y sinonímia del término semilla.

3.1. Obtención del corpus

Un corpus lingüístico es una colección de textos de materiales escritos y/o hablados, seleccionados y ordenados de acuerdo con criterios explícitos, con la finalidad de ser usados como muestra de la lengua. Por otra parte, un corpus computacional es un corpus codificado en una forma estandarizada y homogénea para tareas de recuperación [Sinclair, 1996, Sierra, 2008].

La principal función de los corpus es el estudio del lenguaje humano. Éstos ayudan a mejorar los métodos automáticos de procesamiento del lenguaje y a facilitar el análisis lingüístico.

[1] http://www.graphviz.org/

Algunos aspectos que se deben tomar encuenta para el diseño de un corpus según Torruella and Llisterri (1999), son:

- **Finalidad** concreta para la que servirá.
- **Límites del corpus** que se han de establecer como límites temporales, geográficos, lingüísticos, etc.
- **Captura de los textos y etiquetado**, la captura se puede realizar a través de Internet accesando a una gran cantidad de textos digitalizados de todo tipo, mejor conocidos como corpus informatizados; el etiquetado dependerá del análisis lingüístico que se requiera.

Un criterio muy general para distinguir entre los tipos de corpus según Martí (2003) y Gonzalez et al. (2008) son:

- Corpus orales: Recogen grabaciones o representacciones de grabaciones con alfabeto fonético.
- Corpus escritos: Recogen información sólo de textos.

Existen diversos criterios de clasificación según la función de los parámetros que se utilizarán. Uno de los criterios que mencionan Torruella y Llisterri (1999) es *según la especificidad de los textos que lo componen*, la cual se divide en:

- Corpus general: Recoge textos que reflejen la lengua común en su sentido más amplio.
- Corpus específico: Es un corpus especializado que recoge textos que puedan aportar datos para la descripción de un tipo particular de lengua, es decir, es el conjunto de textos especializados en una área del conocimiento como el Corpus de Ingeniería.

Otro criterio es *según la codificación y la anotación* que clasifica a los corpus atendiendo a las etiquetas descriptivas y analíticas que se han usado en la codificación de los textos como:

- Corpus simple: Es el corpus que ha sido guardado en formato neutro, también llamado plain text y sin etiquetas para ninguno de sus aspectos.

- Corpus codificado o anotado: Es el corpus formado por textos, donde se les ha agregado, manual o automáticamente, «etiquetas declarativas de algunos elementos estructurales de los documentos (indicación de título, de principio de capítulo, de cambio de lengua, etc.) - codificación- o etiquetas analíticas de algunos aspectos lingüísticos (indicación de frase subordinada, de aspectos pragmáticos, etc.).»[Torruella y Llisterri, 1999, pg.~57]

Según Villayandre (2010) razones por las que se usan corpus electrónico es debido a la velocida y precisión del procesamiento automático de los datos, lo que implica la facilidad de manipular y accesar a los datos.

3.2. Extracción de los términos de la taxonomía

Los contextos definitorios se pueden encontrar en textos de cualquier temática para definir los términos usados y contextualizar de mejor forma al lector. La importancia de extraer los contextos definitorios se debe a que éstos proveen las relaciones léxicas que permiten encontrar la jerarquía entre los términos, facilitando de esta forma la generación de las taxonomías.

El método propuesto en el presente trabajo aborda el problema de la extracción automática de hipónimos, hiperónimos y sinónimos a partir de relaciones léxicas que se encuentran en textos no estructurados tomados de la web. La idea principal es usar algunas de las relaciones léxicas propuestas por Ortega (2007) para generar la taxonomía del PLN, además de proponer nuevos patrones léxicos para la extracción de taxones.

El método que se usará es el *basado en patrones*, el cual reconoce los patrones léxicos en un texto. Se eligió este método debido al alto nivel de precisión que presenta comparado con otros métodos mencionados en el capítulo 2 (sin embargo, no está excento de extraer información incorrecta). Otro motivo es que no depende de herramientas como etiquetadores para encontrar los taxones que conformarán la taxonomía, pues no incluyen información morfológica.

Para la extracción de ideas completas se considerarán las oraciones de punto a punto. De las oraciones se selecionarán aquellas que contenga alguna de las relaciones léxicas propuesta en el trabajo de Ortega (2007) y el término semilla dado por el usuario que servirá como término de partida para la realización de la taxonomía.

Cabe señalar que para mejorar el sistema podría realizarse un etiquetado POS[2] sólo en los fragmentos que contengan el término semilla para filtrar los elementos encontrados y así obtener mejores resultados. Sin embargo, no se realiza porque es necesario observar la eficiencia de no usar etiquetados.

Se considera que una pareja de hipónimo-hiperónimo y sinónimo es pertinente si varios patrones la extraen, y de igual manera, un patrón será adecuado mientras mayor número de parejas correctas recupere [Ortega, 2007].

Los patrones léxicos que se usarán son:

1. <hipónimo>es el <hiperónimo>
2. <hipónimo>es un <hiperónimo>
3. <hiperónimo>: <hipónimo>
4. <hiperónimo>(.*) puede definirse <hipónimo>
5. <hipónimo>puede considerarse <hiperónimo>
6. <hipónimo>(.*)se entiende <hiperónimo>
7. <hiperónimo>(.*)se puede <hipónimo>
8. <hipónimo>(.*)se aplica <hiperonimo>
9. <hiperónimo>(.*)se concibe <hipónimo>
10. <hiperónimo>6 palabras[3]: <hipónimo>
11. <sinónimo>o <sinónimo>

3.3. Presentación de la taxonomía

Es importante realizar la representación de la taxonomía, debido a que el uso de estas herramientas o técnicas gráficas permiten la asimilación de la información de forma más rápida y clara.

[2]Por sus siglas en inglés Part-Of-Speech, también conocido como etiquetado gramatical de palabras. «Su finalidad consiste en asociar a cada palabra del corpus una etiqueta representativa de la clase de palabras a la que pertenece [Cózar, 2006, pg.~121]».

[3]Se refiere a que existe como máximo una distancia de 6 palabras a los dos puntos.

Además hay que recordar que «una de las funcionalidades de los sitios web en los que la taxonomía juega un papel protagonista en la búsqueda de información. Los sistemas que permiten buscar contenidos en el entorno web pueden clasificarse en tres grandes tipos: de exploración ("browsing"), de recuperación ("searching") y de filtraje ("filtering")» [Centelles, 2005].

Existen páginas en la web donde se pueden desarrollar diferentes representaciones para los vocabularios controlados (tesauros, ontologías, taxonomías, etc.)

Centelles (2005) muestra algunas opciones para la presentacion de taxonomías corporativa que acontinuación se describen:

- Presentación íntegra, que consta de mostrar sus categorías y las relaciones que la interconectan
- Presentación parcial de la taxonomía original, se eliminan los contenidos a partir de criterios temporales o de uso.
- Reducción de la taxonomía a la relación de equivalencia, de forma que la taxonomía adopta la forma de anillo de sinónimos.
- Reducción de la taxonomía a la relación jerárquica, se reducen los niveles de amplitud y profundidad para ajustar la taxonomía debido a ciertos criterios como capacidad visual, etc.
- «Presentaciones alternativas, como pueden ser la ordenación alfabética de las categorías o las presentaciones arbórea, gráfica y metafórica.»[Centelles, 2005]

Uno de los temas centrales en la IA moderna es la representación del conocimiento. Los métodos que se utilizan son combinaciones de estructuras de datos no lineales, ya que se caracterizan para representar datos con una relación jerárquica entre sus elementos, como lo son los árboles y grafos [Gómez y de~Jesús, 2008, Barceló, 2009].

El árbol es una estructura de datos no lineal, aplicada sobre una colección de elementos u objetos llamados nodos. Además tiene relaciones jerárquicas de parentesco entre los nodos, dando lugar a términos como padre, hijo, hermano, antecesor, suscesor, antepasado, etc.[Gómez y de~Jesús, 2008, Desongles, 2005].

Al nodo princial se le conoce como raíz y los útlimos nodos se llaman nodos terminales u hojas; mientras que los nodos que tienen hijos se llaman nodos no terminales, nodos internos o ramas. [Gómez y de~Jesús, 2008, Desongles, 2005]

Gómez et al. (2008) menciona que los grafos o gráficas son una estructura de datos no lineal, empleados para representar la relación existen entre varios objetos. Además de ser una Teoría Combinatoria muy importante en Matemáticas.

En resumen, la metodología consta de tres etapas: La primera etapa es construir un corpus con la finalidad de obtener la terminología asociada al PLN, es decir, se compilaran aquellos textos que contengan explicitamente el término Procesamiento del Lenguaje Natural o PLN o Lingüística Computacional.

El segundo paso será la extracción de la taxonomía a partir de patrones léxicos ya propuestos en otras investigaciones. Debe recordarse que estos patrones ya definen donde se encontrarán los hipónimos y los hiperónimos.

Por último, la presentación gráfica de la taxonomía se realizará de forma manual de acuerdo a los resultados obtenidos en el paso anterior.

Capítulo 4

Extracción de la taxonomía del PLN

Primeramente en este capítulo se presentará una breve explicación de las herramientas de programación utilizadas para el desarrollo del sistema supervisado para la extracción de taxonomías. Posteriormente, se describirán las pruebas realizadas.

Se realizaron otras pruebas al sistema elaborado para verificar el funcionamiento en la extracción de páginas web, del término semilla y de los patrones léxicos usados. En la extracción de páginas web se probaron desde términos como nanotecnología hasta nombre de actores para verificar la extracción de contenido web. Pero no se presentan debido a que escapan al dominio de la tesis.

4.1. Herramientas de programación utilizadas

En esta sección se presentan una descripción del lenguaje de programación y las herramientas utilizadas para el desarrollo de este trabajo.

4.1.1. Equipos usados

En la tabla 4.1 se describen las características de los equipos usados en la elaboración de la taxonomía semi-automática.

Características	Equipo 1	Equipo 2
Procesador	Atom	Intel Core 2 Quad
Memoria RAM	1GB	4GB
Tipo de sistema	32 bits	64 bits
Sistema operativo	XP y Ubuntu 10.0	Windows Vista

Tabla 4.1: Características de los equipos usados

4.1.2. Python

Python es un lenguaje de programación utilizado para el PLN, simple y rápido [Bird, Klein y Loper, 2009]. Se eligió este lenguaje porque cuenta con funciones básicas predeterminadas para el procesamiento de información lingüística de forma eficiente, a su vez es flexible con la la lectura y escritura de los documentos. Además la ventaja que presenta es que sólo necesita pocas líneas de código para analizar los textos, estas son algunas razones por las cuales es uno de los lenguajes comunmente usados en el procesamiento de textos.

En otras palabras, las ventajas que muestra comparado con otros lenguajes de programación, es que Python es un lenguaje compacto, portable, claro, con una sencilla sintaxis. Este lenguaje de programación permite al programador escribir códigos fáciles de leer y hacer análisis más complejos a los textos.

Otra característica de Python es la compatibilidad con otros lenguajes de programación como: C, C++, Java, .NET; además la ejecución puede realizarse en cualquier sistema operativo como: Windows, Unix/Linux, Mac, etc.

Nota. *La versión 2.7.2. de Python fue utilizada en este proyecto.*

Los módulos que se utilizaron para desarrollar los programas para la extracción de taxonomías se muestran en la tabla 4.2

Módulo	Funcionamiento
re	Para el manejo de expresiones regulares.
os	Para el manejo de archivos.
sys	Para la impresión del resultado de consola en un archivo.
glob	Para el manejo de referencias a archivos.

Tabla 4.2: Módulos de Python usados para la extracción de taxonomías

4.1.3. Segmentador oracional

El segmentador oracional es una herramienta computacional que sirve para dividir un texto por oraciones. El segmentador oracional empleado es el de Munoz y Nagarajan (2001) de la univerdad Illinois, el cual usa un diccionario de abreviaturas en inglés para segmentar las frases del texto. Sin embargo, el diccionario de abreviaturas se cambió por uno en español obtenido de la RAE [1]. Se eligió esta herramienta porque presenta una buena segmentación oracional y con una salida libre de etiquetas.

En particular, esta herramienta segmenta cuando se presenta un punto y aparte seguido de una létra mayúscula o cuando se encuentra un punto y seguido y la siguiente palabra comienza con mayúscula.

Nota. *El corpus generado se procesa con el segmentador oracional mencionado.*

Con una prueba realizada al segmentador oracional de Munoz y Nagarajan (2001) se observó que su precisión es aproximadamente del 80 %.

4.1.4. Solid Converter

Es un software que permite convertir archivos de formato PDF [2] a DOC [3], a hojas de cálculo Excel, txt y a muchos otros formatos [documents, 2010]. Se eligió este software por ser una herramienta gratutita para la conversión de archivos en diferentes formatos.

[1] http://lema.rae.es/dpd/apendices/apendice2.html
[2] Son las siglas en inglés de Portable Document Format, formato de documento portátil.
[3] El formato de archivo utilizado por el procesador de texto Microsoft Word.

Nota. *Se emplea Solid Convert para la conversión de los documentos PDF a TXT del corpus generado.*

4.1.5. Graphviz

Graphviz es un software de vizualización de gráficos gratuito y de código abierto. Sirve para representar información estructural con diagramas gráficos, abstractos y de redes. Pricipalmente tiene aplicaciones en redes de datos, bioinformática, ingeniería de software, base de datos, diseño de páginas web, aprendizaje de máquinas y en las interfaces visuales para otros dominios técnicos [Bilgin, Ellson, Gansner, Hu, North, Koren, Dobkin, Dwyer, Koutsofios, Lilly, Low, Mocenigo, Scheerder, Woodhull y Caldwell, 1999].

Nota. *Se utliza para presentar graficamente la taxonomía del PLN.*

4.1.6. ECODE

El Extractor de Contextos Definitorios «ECODE» es un sistema basado en reglas lingüísticas que incorpora diferentes procesos para la extracción de contextos definitorios a partir de patrones verbales en textos de especialidad. El resultado entregado es un conjunto de contextos definitorios que son almacenados en nuevo archivo [Alarcón, 2009].

En otras palabras, el sistema ECODE extrae automáticamente contextos definititorios de archivos TXT, PDF, HTML, XML, DOC o de alguna URL proporcionada por el usuario al sistema.

El algoritmo de ECODE se implementó en PERL[4] y consta de nueve módulos según explica Alarcón (2009). En la fig. 4.1 puede observarse su arquitectura general. Dentro de las funciones específicas que realiza son:

1. Búsqueda de un término en específico dentro del documento.
2. Etiquetado POS.
3. Segmentación oracional.
4. Lematización[5] de los CDs encontrados en un texto.

[4]Es un lenguaje de programación generalmente usado en el PLN
[5]Lematizar es reducir al infinitivo masculino las palabras

Los resultados que presenta el ECODE son:

- Número total de CDs
- Términos totales encontrados
- Número de CDs que contiene el término buscado
- Tiempo de procesamiento

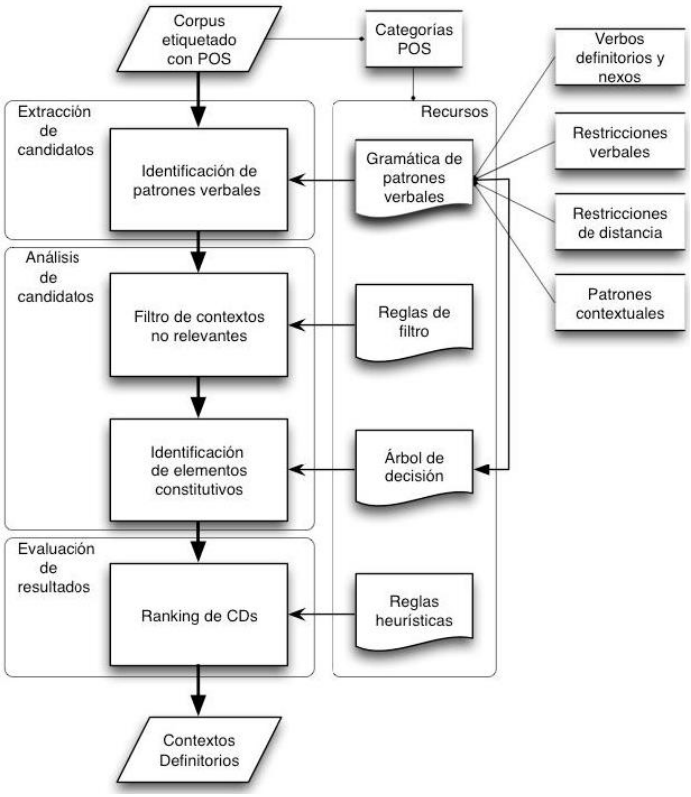

Figura 4.1: ECODE. Fuente: Alarcón. (2009: pp. 143)

ECODE presenta una precisión = *38.953 %* y Exhaustividad = *88.157 %* [Vieyra, 2011].

4.2. Proceso de la extracción de taxonomías

En la fig. 4.2 se muestra el proceso realizado para la extracción taxonomías de forma supervisada.

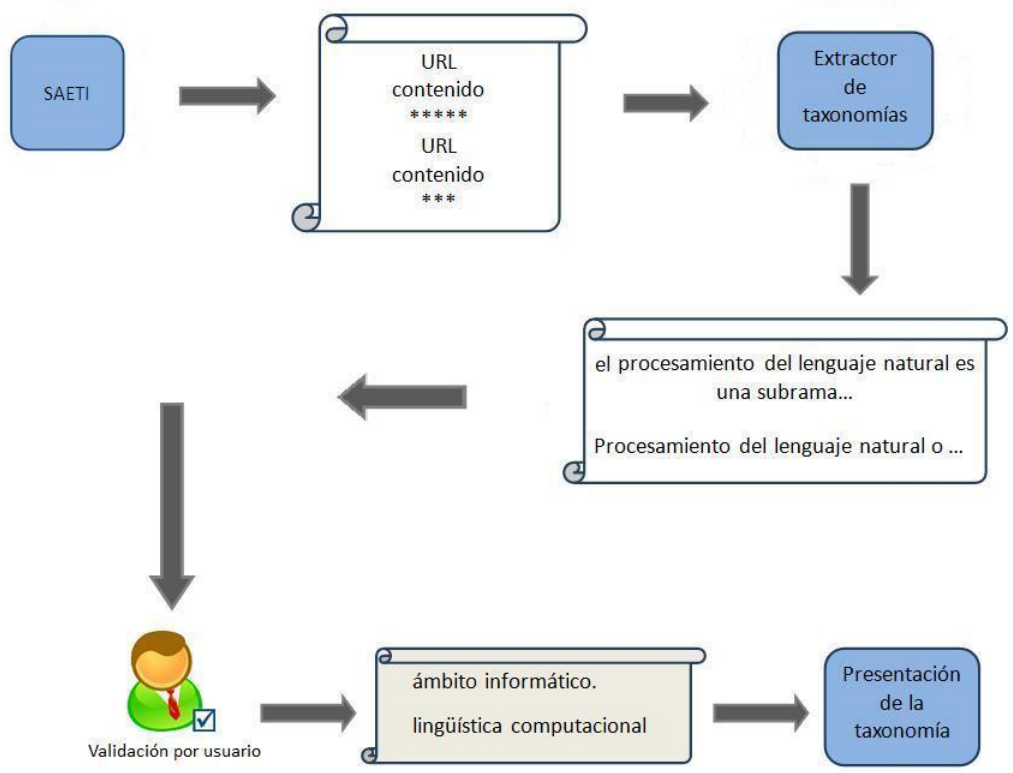

Figura 4.2: Proceso supervisado en la extracción de taxonomías

El proceso para la extracción de taxonomías consiste de dos módulos: el extractor de páginas y el extractor de taxonomías. Cada uno de los módulos genera archivos de resultados respectivamente.

El primer módulo, el Sistema Automático de Extracción de Textos de Internet «SAETI» fig.4.3 es un extractor de páginas web, desarrollado en colaboración con el Ing. Josué Antonio Careaga Moya, M. Octavio Augusto Sánchez Velázquez y la Dra. Fernanda López Escobedo, para la compilación automática de textos web que contengan un término, palabra o frase determinada por el usuario.

El módulo «SAETI» a su vez consta de los siguientes módulos:

1. Extracctor de snippets[6]: Realiza las consultas automáticas a Google.
2. Extractor de URLs: Extrae las URLs de las páginas web que contienen el término, palabra o frase determinada por el usuario.
3. Extractor de contenido: Extrae el contenido de cada página web, filtrando las etiquetas HTML.

El primer módulo del «SAETI» puede generar hasta 800 consultas. El resultado obtenido de cada búsqueda es una lista de ligas a las páginas web y en cada liga muestra un párrafo que contiene la palabra buscada (Snippet). El listado se guarda en un archivo de texto que servirá de entrada para el siguiente módulo.

Ejemplo 12 (Resultado del primer módulo del SAETI)**.** *http://www.esi.uem.es/~jmgomez/pln/index.html*

*< b >Procesamiento del Lenguaje Natural< /b > El < b >Procesamiento del Lenguaje Natural(PLN) es la disciplina encargada de
producir sistemas informáticos que posibiliten dicha comunicación, por medio de
la ...
 </cite>www.esi.uem.es/ jmgomez/pln/index.html </cite>/url?q=http ://nlp.uned.es/&sa=U&ei=PQJhT_HHHoKg2gWiw9GUCA&ved=0 CCoQFjAF&usg=AFQjCNGHadvuS15WWm4kWM4_0iea5MhYOw Natural Language Processing and Information Retrieval Group at ...Natural Language Group at the Spanish National Distance University (UNED).
Research on natural language processing applied to information access, ...
 </cite>nlp.uned.es/</cite>*

El segundo módulo del «SAETI» filtra a partir de expresiones regulares el archivo obtenido en el primer módulo y entrega como resultado un archivo de salida que sólo contiene la lista de URLs. Este archivo servirá como entrada para el siguiente módulo.

[6]En SEO «Searching Engine Optimization (Motor de búsqueda optimizada)» el snippet es el resultado mostrado por los buscadores en las consultas realizadas. Se encuentra formado por la URL y un fragmento que contiene las palabras de la consulta.

Ejemplo 13 (Resultado del segundo módulo del SAETI). *http://www.sepln.org/*
http://www.lsi.upc.edu/
http://www.esi.uem.es/
http://nlp.uned.es/
...

El último módulo del «SAETI» extrae de cada una de las ligas del archivo de resultados del segundo módulo el contenido de la página web filtrado, es decir, sin código HTML. Para realizar la separación de información entre páginas web se le agregaron dos etiquetas:

1. La primera etiqueta indica la url del cual fue extraído el contenido.
2. La segunda hace la división entre cada página web obtenida. El archivo generado en este módulo es el resultado final del SAETI.

Ejemplo 14 (Resultado del último módulo del SAETI). *procesamientolenguajerecuperacion.50webs.org/*

El procesamiento del lenguaje natural es una subrama de la inteligencia artificial y de la lingüística. También se suele referir a esta rama de la informática de forma abreviada como PLN o NLP, por sus siglas en inglés Natural Language Processing.
...

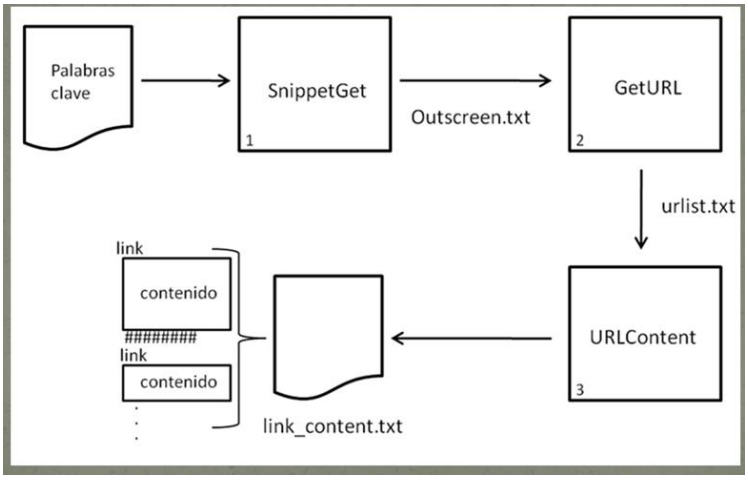

Figura 4.3: Arquitectura del SAETI

La forma estándar de usar el programa es ingresar un término, aunqe también puede ser usado por módulos, ésto es ingresar un archivo que contenga una lista de snippets y se obtendrá una lista de urls, o ingresar un archivo con una lista de urls y se obtendrá una lista de contenidos de las urls, como se puede observar en la fig. 4.4.

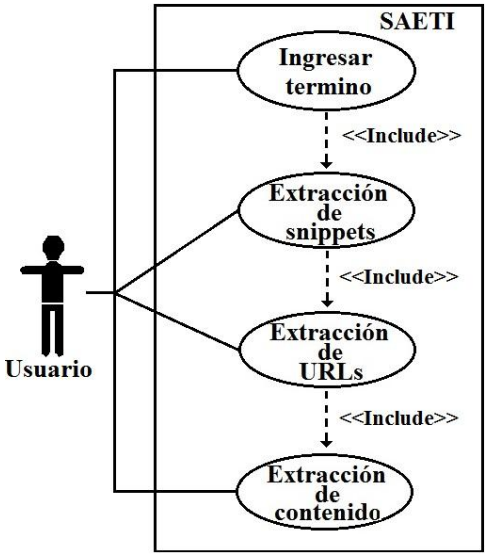

Figura 4.4: Diagrama de casos de uso del SAETI

Escenario: Obtención de corpus

Actor: Usuario del sistema

Descripción: El usario del sistema llevará acabo la obtención de un corpus web a partir de un término semilla, es decir, el usario escribirá en el sistema el término del cual se realizará una compilación de los textos que contengan dicho término.

Primero el usuario debe escribir en el sistema un término para realizar la consulta en la web.

Segundo el sistema obtiene las snippets relacionadas al término escrito por el usuario.

Después el sistema procesa la información para obtener una lista de URLs relacionadas al término.

Posteriormente el sistema copia y guarda el contenido de cada URL en un archivo de texto.

Excepciones: Si la URL que se arrojó en la búsqueda del término es de tipo PDF, DOCX o PPT no se extraerá su contenido debido a que la mayoría de estas extensiones con contenido científico se encuentran encriptados.

El segundo módulo, el «Extractor de Taxonomías» fig. 4.5 cuenta con los siguientes cuatro módulos:

1. Filtro: Limpia el archivo de entrada generado por módulo del «SAETI», es decir, elimina todo aquel texto que no pertenezca a la página web.
2. Segmentador oracional: Divide por oraciones el texto.
3. Extractor de frases: Extrae las oraciones que contienen el término determinado por el usuario, llamado en esta tesis término semilla.
4. Clasificador de frases obtenidas: Clasifica las oraciones del corpus que contienen el termino semilla de acuerdo al patrón léxico que presenten.

El primer módulo del extractor de taxonomías recibe como entrada el archivo resultante generado por el SAETI el cual es filtrado para eliminar las URLs colocadas para segmentar por documentos. El segundo módulo segmenta oracionalmente el archivo ya filtrado del paso anterior, dando como resultado un archivo que sólo contiene el texto web.

El tercer módulo extrae las oraciones que contienen el término semilla del archivo resultante del segundo módulo. El último módulo clasifica todas las oraciones obtenidas del paso anterior de acuerdo a los patrones léxicos programados. El resultado obtenido en éste módulo es un archivo con un listado de fragmentos de oraciones que contienen los hiperónimos, hipónimos y sinónimos del término semilla. Finalmente, un experto selecionará y validará de estos fragmentos los términos que encuentre.

En la fig. 4.5 se muestra la arquitectura del extractor de taxonomías. Los bloques en azul son los módulos de la herramienta, los pergaminos blancos son los archivos generados a la salida de los módulos y el pergamino morado es el último archivo generado del extractor de taxonomías, el cual muestra el resultado obtenido por este módulo.

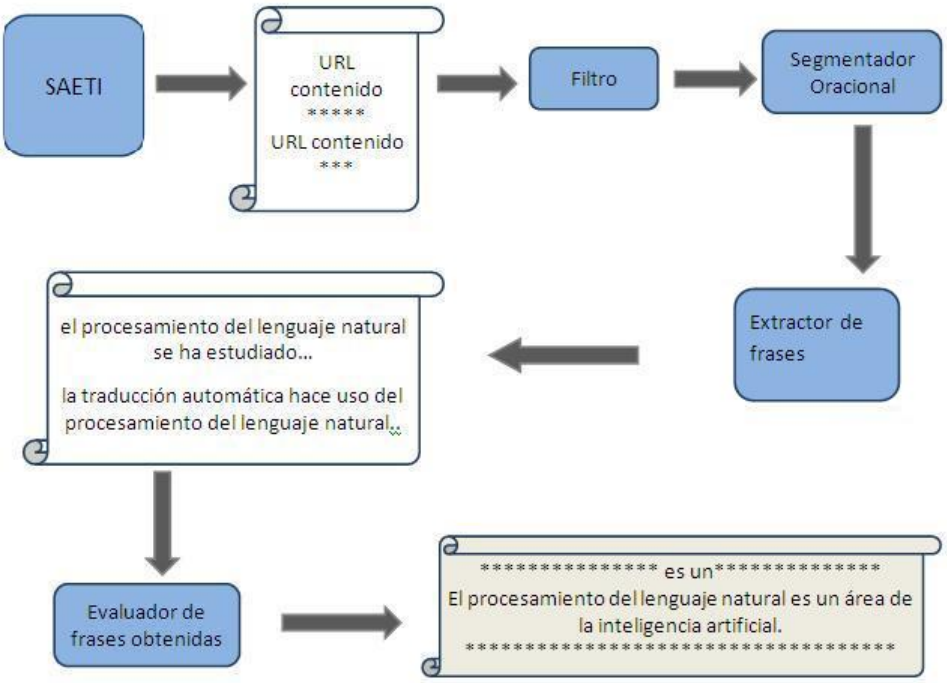

Figura 4.5: Arquitectura del extractor de taxonomías

La forma de usar el programa es ingresar un corpus para obtener la taxonomía. El sistema permite utilizar los módulos por separado, como se puede observar en la fig. 4.6.

Escenario: Extracción de una taxonomía

Actor: SAETI

Descripción: A partir del corpus obtenido por el SAETI se obtendrá la terminología asociada al término escrito por el usario para generar una taxonomía.

Primero el sistema elimina las etiquetas del resultado generado por el SAETI.

Segundo el sistema segmenta por oraciones el corpus.

Tercero el sistema procesa la información para obtener una lista de oraciones que contenga el término escrito por el usuario en el SAETI.

Cuarto el sistema clasifica las oraciones que contienen el término de acuerdo a su relación léxica y las guarda en un archivo de texto.

Finalmente un experto valida los términos clasificados.

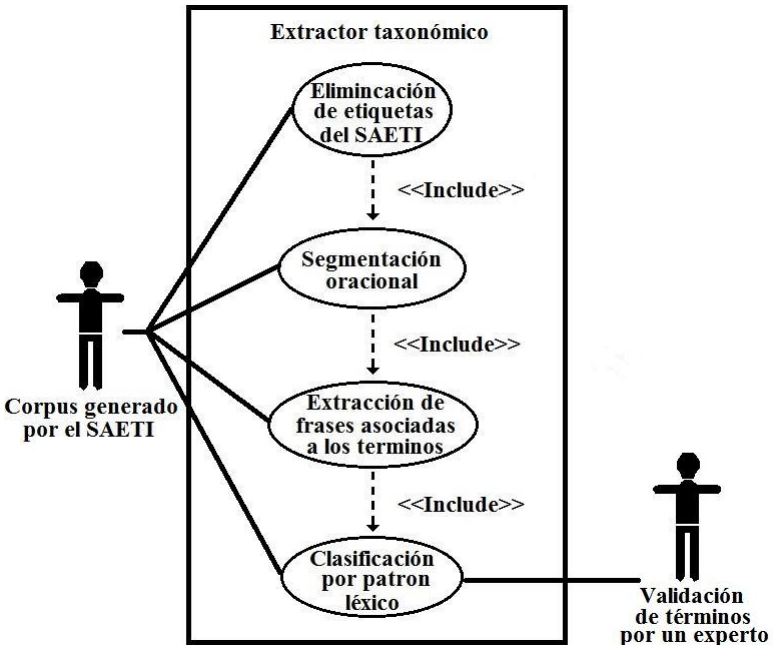

Figura 4.6: Diagrama de casos de uso del Extractor de Taxonomías

Después de emplear los módulos anteriores se puede graficar para ejemplificar de mejor forma la taxonomía obtenida.

Con ayuda de Graphviz se genera una gráfica en forma de árbol a partir de aquellas palabras o términos dados por el usuario. Para darle la jerarquía a cada término deberán ser escritos uno por línea y posteriormente serán alineados de acuerdo a su jerarquía.

```
Procesamiento del lenguaje natural
        Corpus
                Etiquetado Automático
                        Etiquetado POS
```

Figura 4.7: Jerarquía en Graphviz

La idea del extractor de taxonomías es generar una taxonomía a partir de un término dado de cualquier área del conocimiento, a través de recuperación y extracción de páginas web. Posteriormente se extraen aquellas frases que contienen los patrones léxicos que permiten encontrar los taxones relacionados al término semilla. Por último, se evalua manualmente si la palabra encontrada en la frase es o no un término.

4.3. Textos de PLN

Para la extracción de los términos, en nuestro caso llamados taxones, que conforman la taxonomía en PLN se compilaron tres corpus, en los cuales se realizaron pruebas diferentes usando el método mencionado en el tercer capítulo.

La finalidad de la compilación del corpus es obtener y capturar suficientes textos actualizados de la web que sirvan como base para la extracción de la taxonomía de PLN en español.

El primer corpus elaborado fue de forma manual, realizando una búsqueda y elección de textos sobre PLN en *Google académico*, donde apareciera el término Procesamiento del Lenguaje Natural, Lingüística Computacional, traducción automática o resumen automático, etc. Estas áreas se buscaron debido a que son los más conocidos a nivel mundial.

La razón de realizar la búsqueda en *Google académico* es debido a la validez que presentan los documentos recuperados en esta sección del buscador, ya que son tesis, artículos de revistas científicas o capítulos de libros. Lo anterior facilita verificar la procedencia de la información y así presentar los términos usados adecuadamente.

Entre las dificultades presentadas durante la generación del corpus, estuvo la escasez de textos en español acerca del Procesamiento del Lenguaje Natural o Lingüística Computacional y sus diversas ramas de estudio. Además de encontrarse con varios textos encriptados, imposibilitando el procesamiento de los mismos.

Los criterios usados en la compilación de los tres corpus fueron:

1. Se trata de un *corpus escrito* debido a que se extrajeron sólo textos de la web.
2. Es un *corpus específico* porque los textos usados pertenecen a una área del conocimiento como lo es el procesamiento del lenguaje natural
3. Es un *corpus simple* ya que el texto extraído se guardó en un documento sin formato.
4. También es un *corpus codificado o anotado* porque se etiquetaron las relaciones léxicas encontradas en los textos para hacer más visibles los taxones encontrados.

En la tabla 4.3 se describen las características del corpus obtenido manualmente.

Característica	Descripción
Búsqueda realizada	Temas asociados al procesamiento del lenguaje natural.
Formato de textos	Textos digitalizados.
Extensión de los archivos	PDF y DOC.
Tipo de textos	Académicos como tesis de licenciatura, maestría y doctorado, publicaciones en revistas, pósters, libros y actas en congresos.
Longitud de los textos	Variable
Temática	Textos referentes explicítamente al procesamiento del lenguaje natural, PLN o lingüística computacional, como traducción automática, extracción de CDs, resumen automático, etc.
Idioma	Español
Total de textos	80: De los cuales 40 artículos de revistas, 6 actas de congresos, 5 capítulos en libros, 5 tesis de doctorado, 15 tesis de maestría, 7 tesis de licenciatura y 2 informes de congresos.
Tiempo de elaboración	14 semanas (2.5 meses).
Fuente	Google Académico
Codificación	UTF-8

Tabla 4.3: Primer corpus

En la tabla 4.4 se muestra el segundo corpus que se generó de forma automática a partir de textos web asociados a un término semilla propuesto por el usuario. En el caso particular de esta tesis son los términos de Procesamiento del Lenguaje Natural y Lingüística Computacional seguidos de alguna de las relaciones léxicas propuestas por Ortega (2007).

En este caso no se presentaron buenos resultados en la recuperación de sus respectivos documentos, ya que Google no tomaba en cuenta los signos de puntuación para la recuperación de información. Se encontraron, por ejemplo, el *procesamiento del lenguaje natural, es una ciencia...*; razón por la cual el análisis no fue satisfactorio.

Característica	Descripción
Búsqueda realizada	Procesamiento del lenguaje natural, Lingüística computacional, Resumen Automático y Traducción Automática.
Formato de textos	Textos digitalizados.
Extensión de los archivos	PDF, DOCX, PPT
Tipo de textos	Académicos
Longitud de los textos	Variable
Temática	Textos referentes explicítamente al procesamiento del lenguaje natural, PLN o lingüística computacional, como traducción automática y resumen automático
Idioma	Español
Total de textos	100 documentos
Tiempo de elaboración	1 semana
Fuente	Google Académico
Codificación	UTF-8

Tabla 4.4: Segundo corpus

El tercer corpus fue generado también de forma automática. Las dificultades que se presentaron durante la recuperación de los archivos fue que se encontraban en diferentes codificaciones, es decir, la mayoría de las veces los acentos no se presentaban de una forma estándar.

Además los textos con extensiones PDF, DOCX y PPT se dejaron de tomar en cuenta ya que la mayoría de las veces este tipo de documentos se encontraban encriptados, imposibilitando la extracción de la información.

En la tabla 4.5 se describen las características del corpus obtenido automaticamente.

Característica	Descripción
Búsqueda realizada	procesamiento del lenguaje natural y lingüística computacional seguidas o precedidas de alguna de las relaciones léxicas propuestas por Ortega (2007).
Formato de textos	Textos digitalizados.
Extensión de archivos	HTML, PDF y DOC.
Buscador	Google.
Tipo de textos	Páginas web que presentaran el término procesamiento del lenguaje natural, PLN o lingüística computacional.
Longitud de los textos	Variable
Temática	Procesamiento del lenguaje natural y PLN.
Idioma	Español
Total de textos textos	87 URLs de lingüística computacional, 59 URLs de procesamiento del lenguaje natural o PLN y 89 URLs de procesamiento de lenguaje natural, dando un total de 235 URLs. Cada URL se tomó como un texto.
Tiempo de elaboración	3 semanas.
Fuente	Google Académico
Codificación	UTF-8

Tabla 4.5: Tercer corpus

4.4. Contextos definitorios obtenidos

Se empleó ECODE en el análsis del corpus elaborado manualmente con el fin de obtener los contextos definitorios asociados al PLN y así recuperar los fragmentos que presentaran las relaciones léxicas para generar la taxonomía.

Pero la primera dificultad que se presentó fue la encriptación de varios textos del corpus, principalmente en aquellos archivos con extensión PDF y DOC, imposibilitando la extracción de sus respectivos CDs. Otro problema que se presentó en este tipo de documentos fue la segmentación no adecuada por oraciones, lo cual no permitió encontrar todos los CDs en los textos.

Para resolver el problema de la segmentación en los documentos se utilizó el programa Solid Converter para convertir archivos PDF a TXT y posteriormente volverlos a analizar con el ECODE.

Sin embargo, durante el análisis manual de los CDs obtenidos se observó que éstos no eran satisfactorios debido a que no se obtuvieron suficientes CDs bien formados, ya que la segmentación no se realizaba adecuadamente en algunos textos. Por esta razón se decidió realizar otra prueba con otro corpus.

En la segunda prueba, se empleó el segundo corpus generado con el término semilla seguido o precedido de un patrón léxico para obtener sus respectivos términos asociados, esto se realizó para obtener mayores resultados de cada patrón.

Sin embargo, después de la extracción de la información se observó la aparición de signos de puntuación intermedios entre los patrones lingüísticos recuperados y en algunos casos, la existencia de un salto de línea donde se definia el término. Estas características que presentaron los textos imposibilitaron la correcta extracción de los términos asociados, razón por la cual no se tomó en cuenta esta prueba.

Ejemplo 15 (Frases extraídas en la segunda prueba)**.** *el procesamiento del lenguaje natural, es un campo de...*
el procesamiento del lenguaje natural es un,área...

En la tercera prueba se usó el tercer corpus elaborado por medio del SAETI ya mencionado anteriormente. En este corpus se buscaron los términos de Procesamiento del Lenguaje Natural, Lingüística Computacional (por ser un sinónimo del término semilla) y PLN (por ser las siglas del término semilla).

Las búsquedas de los términos mencionados en el párrafo anterior se realizaron por separado, dando finalmente como resultado tres documentos que contienen su información respectiva.

Capítulo 5

Taxonomía generada semi-automáticamente

En este capítulo se presentan los resultados obtenidos por el extractor de taxonomías para el tercer corpus generado a partir de la web para los términos semilla de Procesamiento del Lenguaje Natural, PLN y Lingüística Computacional. Debe recordarse que un taxón será encontrado por el extractor de taxonomías cuando cumpla con los parámetros mencionados. Además la validación de cada uno de los taxones la realizará un experto.

Son importantes las relaciones léxicas en las frases extraídas que contienen el término semilla debido a que se obtiene la jerarquía de los taxones encontrados para formar la taxonomía.

5.1. Agrupamiento de términos obtenidos

En el agrupamiento de términos se muestran las listas de taxones obtenidos para los términos: Procesamiento del Lenguaje Natural, PLN y Lingüística Computacional, ordenados según su jerárquía (hipónimo, hiperónimo o sinónimo). Se agregó el término de PLN en la extracción de frases porque son las siglas del término principal. Sin embargo, no se realizó la búsqueda específicamente con este término porque no generaba resultados relacionados con esta área de conocimiento.

El agrupamiento de términos obtenidos del tercer corpus por el extractor de taxonómico se muestra a continuación.

HIPERÓNIMOS

1. Conocimiento científico
2. Ciencia
3. Disciplina
4. Lingüística
5. Informática
6. Análisis de lenguaje humano
7. Normalización lingüística
8. Ingeniería Informática
9. Filosofía
10. Matemáticas
11. Psicología
12. Ámbito informático
13. Inteligencia Artificial
14. Lenguaje humano
15. Lenguaje de programación
16. Lenguaje natural
17. Estadística
18. Enfoque lingüístico
19. Área de vanguardia

HIPÓNIMOS

1. Traducción automática
2. Ingeniería de la lengua
3. Reconocimiento del habla
4. Corpus lingüístico asistido por ordenador
5. Corrección ortográfica de textos
6. Recuperación de información
7. Interfaces en lenguaje natural
8. Herramienta para la recuperación de información
9. Herramienta para la organización de la información
10. Lenguajes de programación declarativos
11. Prototipo de tractor catalán/castellano
12. Llull

SINÓNIMOS

1. Machine learning
2. NLP
3. Natural Language Processing
4. Lingüística computacional
5. Ingeniería lingüística

5.2. Estadísticas obtenidas

El número total de frases del tercer corpus es de 13,024. De las cuales sólo 505 contienen alguno de los términos semilla (Procesamiento del Lenguaje natural, PLN o Lingüística Computacional). Entonces 12,519 frases no contienen alguno de estos términos, por lo que fueron descartadas. El tiempo de elaboración de la taxonomía del PLN fue de 5 días.

En la fig. 5.1 se muestra el porcentaje de frases descartadas y las frases que fueron analizadas con mayor profundidad en este trabajo.

Figura 5.1: Análisis general de las frases analizadas

De las 505 frases que contenían los términos semilla, 65 frases cumplieron con alguno de los patrónes léxicos, de las cuales sólo 41 proporcionaron información relevante. 24 fueron clasificadas dentro de los patrones programados que no daban información relevante. Además 36 frases que sólo contenían el término semilla pero no entraban dentro de la clasificación de los patrones programados dieron información relevante y 404 fueron descartadas por no tener alguna de las relaciones léxicas mencionadas anteriormente.

En la fig. 5.2 se muestran los resultados obtenidos de las 505 frases que contenían el término semilla.

Figura 5.2: Análisis general de las frases encontradas

Las 24 frases que no proporcionaron información relevante pero que cumplían con los patrones léxicos programados, se debieron a tres casos generales:

En el primer caso el patrón léxico «es el» se encontraba a la izquierda del término semilla; generalmente hacía referencia a preguntas como ¿qué es el procesamiento de lenguaje natural?.

El segundo caso se debió a no tomar en cuenta un espacio antes del término semilla para la extracción de éste. Este caso se presentó específicamente en la extracción de las siglas PLN. Se observó que en muchos casos en la web se suelen escribir palabras juntas sin darles sus respectivos espacios(errores humanos), pero en algunos casos se pueden extraer términos que finalicen con el patrón semilla dado, sepln (sociedad española para el procesamiento del lenguaje natural).

El tercer y último caso se dió durante la extracción de las frases que contenían dos puntos; estas frases no siempre contenían los hipónimos o hiperónimos en la misma línea, por lo que no se aseguraba que la frase extraída se encontrara completa.

En la tabla 5.1 se muestran los patrones programados para generar la taxonomía de PLN y la frecuencia con la que aparecieron en las 41 frases que contenían el término semilla.

Patrón léxico	**Frecuencia**
<hipónimo>es el <hiperónimo>	2
<hipónimo>es un <hiperónimo>	9
<hiperónimo>: <hipónimo>	8
<hiperónimo>(.*) puede definirse <hipónimo>	2
<hipónimo>puede considerarse <hiperónimo>	2
<hipónimo>(.*)se entiende <hiperónimo>	2
<hiperónimo>(.*)se puede <hipónimo>	1
<hipónimo>(.*)se aplica <hiperonimo>	5
<hiperónimo>(.*)se concibe <hipónimo>	1
<hiperónimo>6 palabras: <hipónimo>	3
<sinónimo>o <sinónimo>	6

Tabla 5.1: Frecuencia de aparición de los patrones en el corpus web

De los 41 taxones validados, sólo 36 términos no se encontraban repetidos, dando un total de 19 hiperónimos, 12 hipónimos y 5 sinónimos.

La precisión obtenida del programa extractor de taxonomías fue de 63.07 %, con un cobertura de 53.24 %.

5.3. Resultados manuales y semi-automáticos

En esta sección se mostrarán las taxonomías obtenidas de forma manual y de forma semi-automática del término Procesamiento del Lenguaje Natural, para compararlas entre ellas y observar los términos encontrados en cada una de ellas.

5.3.1. Resultados obtenidos manualmente

Para generar la taxonomía manual se extrajeron de 30 URLs las frases que contenían los términos: Lingüística Computacional, Procesamiento del Lenguaje Natural y PLN. De cada una de las frases encontradas se extrajeron los términos asociados y se jerarquizaron para generar la taxonomía. El tiempo de elaboración fue de 20 días; 79 taxones fueron encontrados para este trabajo. Los resultados se muestra en las listas siguientes.

Se espera obtener algunos de los taxones de la siguientes listas de resultados de la prueba manual en la prueba supervisada.

HIPERÓNIMOS

1. Lingüística
2. Lingüística aplicada
3. Filosofía
4. Psicología
5. Inteligencia Artificial
6. Ciencias
7. Matemáticas
8. Informática
9. Computación
10. Ingenería de telecomunicaciones
11. Sistemas de aprendizaje
12. Sistemas multimedias
13. Entornos de iconos
14. Programación interactiva
15. Realidad Virtual
16. Hipertexto
17. Sonido
18. Aprendizaje estadístico automático
19. Álgebra Lineal
20. Estadística
21. Estadística Bayesiana
22. Ingeniería en Computación
23. Ingeniería lingüística

HIPÓNIMOS

1. Comunicación máquina-humano
2. Simulación de mecanismos de comunicación
3. Comprensión del lenguaje natural en textos y expresiones habladas
4. Reconocimiento del lenguaje humano
5. Análisis morfológico automático
6. Stemming
7. Tokenizing
8. Etiquetadores morfosintácticos
9. Análisis sintáctico automático
10. Analizadores sintácticos
11. Análisis semántico automático
12. Etiquetado semántico
13. Análisis prágmatico automático
14. Generación de frases automáticas

15. Procesamiento automático de la información
16. Reconocimiento de texto manuscrito
17. Corrección de textos
18. Reconocimiento de entidades
19. Traducción
20. Traducción automática
21. Traducción asistida (TAO)
22. Corpus
23. Generación automática de corpus
24. Etiquetado automático o semi-automático
25. Etiquetado POS
26. Ontologías
27. Anotación lingüística basada en ontologías
28. Localización de ontologías
29. Extracción de información
30. Clasificación de documentos
31. Resumen
32. Resumen automático
33. Resumen por extracción
34. Tutores inteligentes
35. Minería de datos
36. Desambiguación de significados
37. Desambiguación léxica
38. Modelos de Markov
39. Análisis de opinión en blogs
40. Sistemas de recomendación
41. Georeferenciación automática de contenidos
42. GIR
43. Detección de plagio
44. Recuperación de la información
45. Web semántica
46. Búsqueda de documentos
47. Recuperación de contenidos multimedia
48. Síntesis del habla
49. Minería de textos
50. Sistemas inteligentes para la educación y el entrenamiento
51. Reconocimiento automático del habla
52. Reconocimiento de voz
53. Síntesis de voz
54. Evaluación de sistemas de procesamiento del lenguaje natural
55. Respuestas a preguntas

SINÓNIMOS

1. Lingüística computacional

En las figuras 5.3 , 5.4 y 5.5 se muestra la represetación gráfica de la taxonomía generada manualmente.

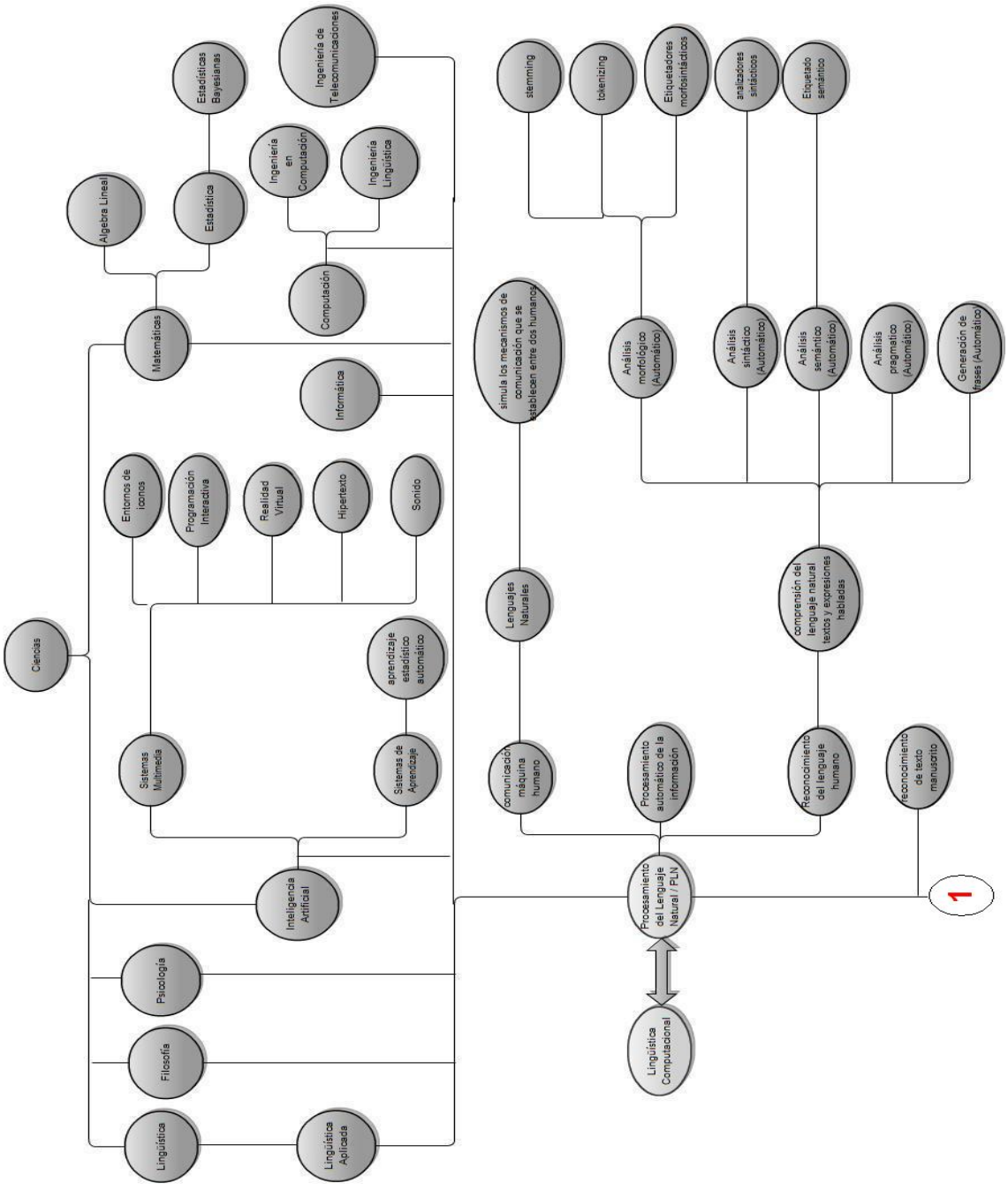

Figura 5.3: Resultados manuales parte 1

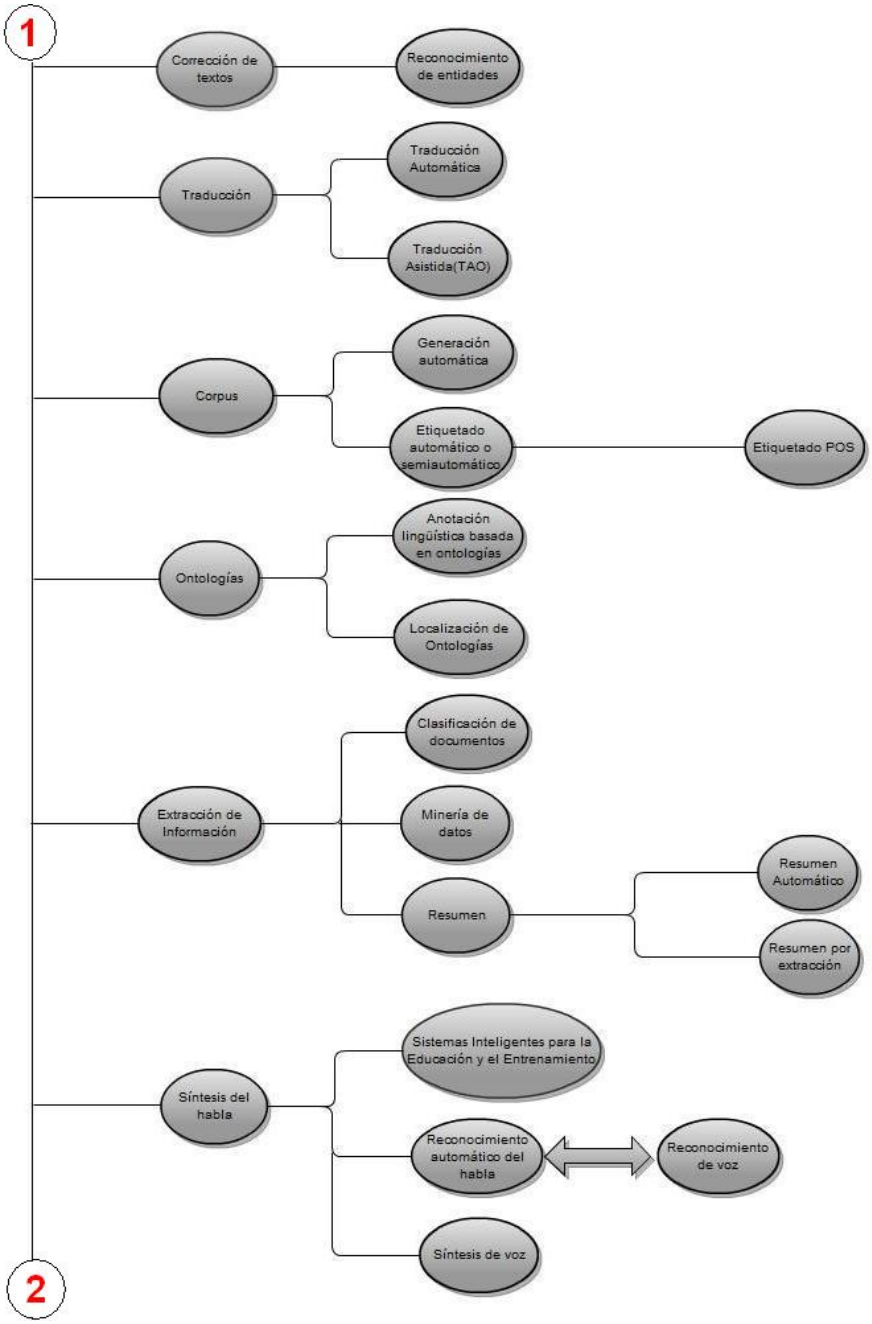

Figura 5.4: Resultados manuales parte 2

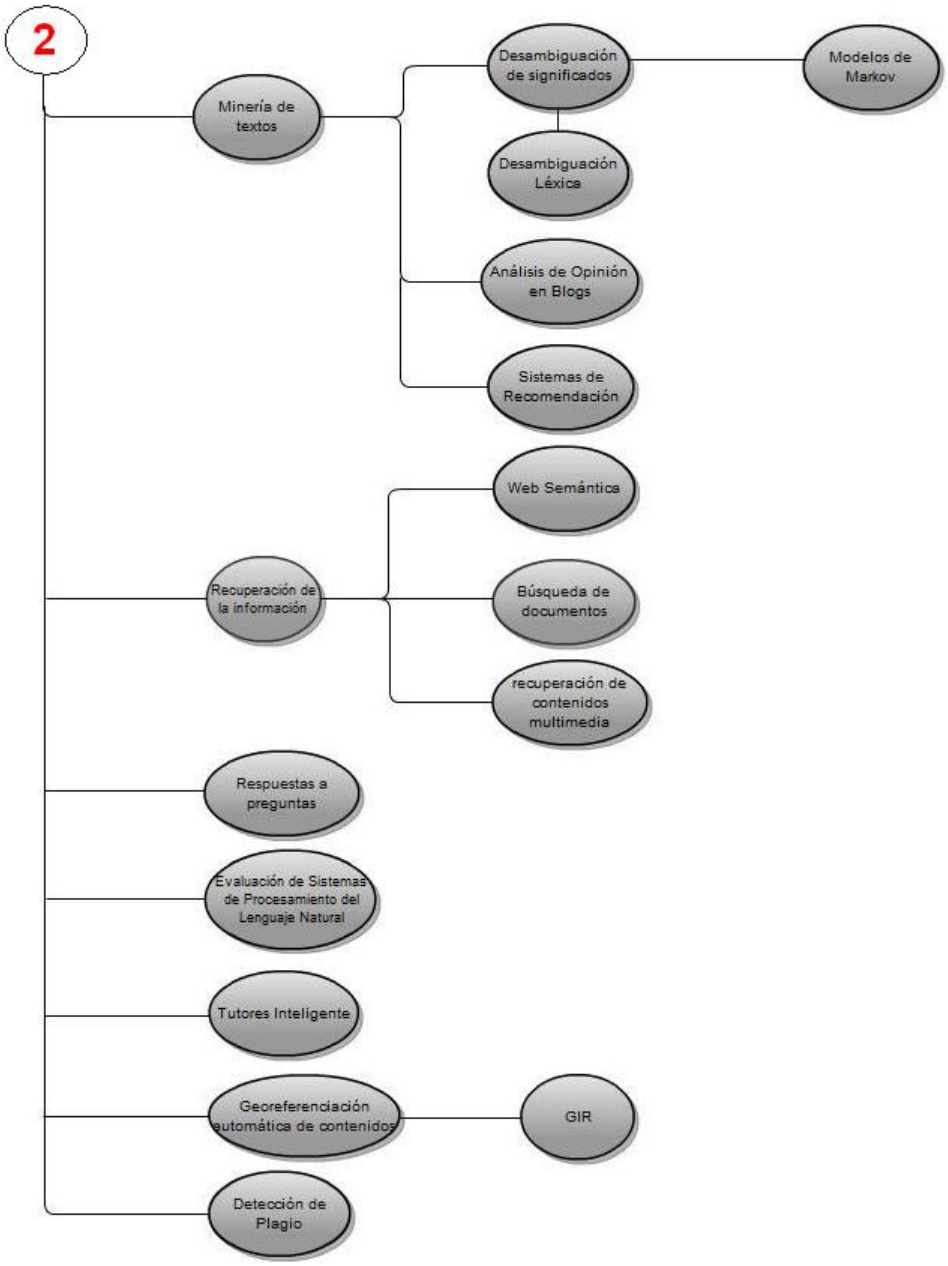

Figura 5.5: Resultados manuales parte 3

5.3.2. Resultados obtenidos semi-auotmáticamente

Con los resultados obtenidos por el extractor taxonómico del tercer corpus se elaboró la representación gráfica de la taxonomía obtenida semi-automáticamente del PLN fig. 5.6, 5.7 y 5.8.

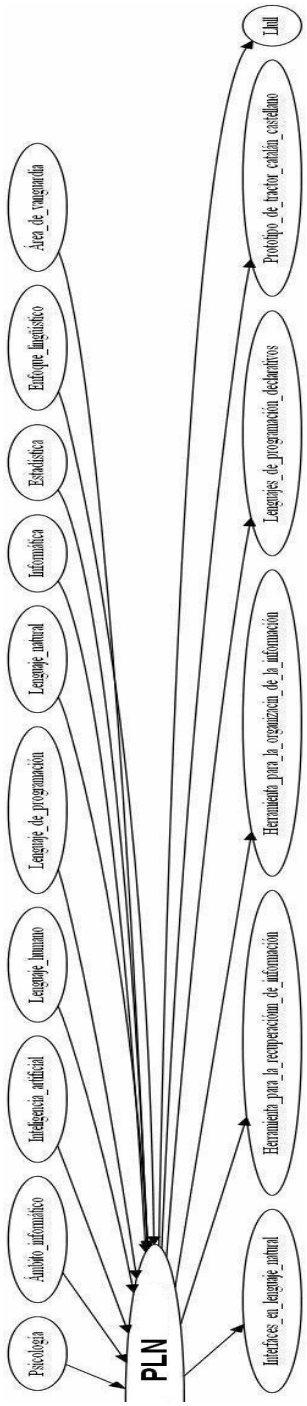

Figura 5.6: Hipónimos e Hiperónimos de PLN 1

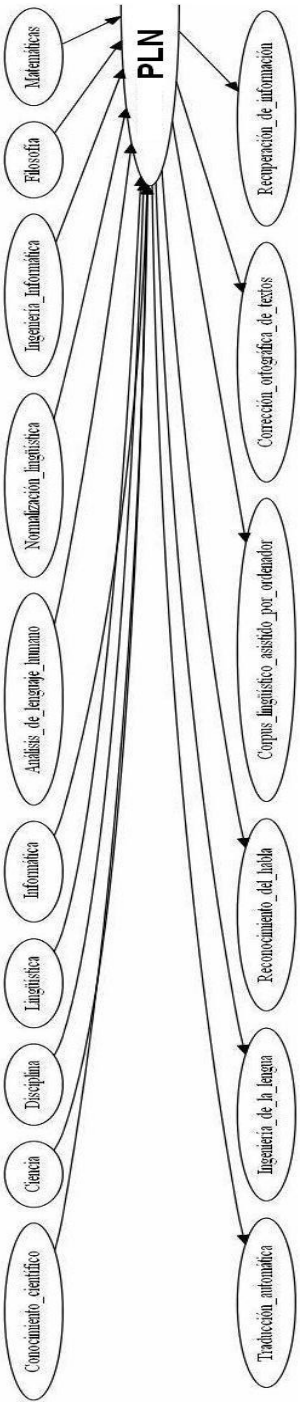

Figura 5.7: Hipónimos e Hiperónimos de PLN 2

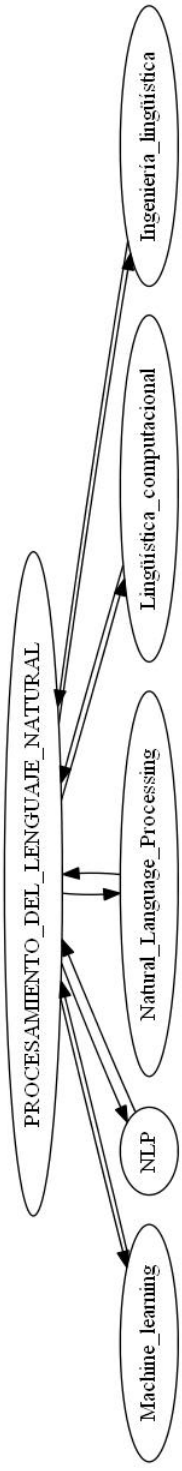

Figura 5.8: Sinónimos del PLN

5.3.3. Comparación de los resultados esperados y obtenidos

En la tabla 5.2 se muestra los resultados tanto manuales como los semi-automáticos. Se puede observar que el tiempo de elaboración de la taxonomía es menor empleando el extractor taxonómico, siendo que sólo se utilizó el 25 % del tiempo que se llevo para elaborar la taxonomía manual.

Sin embargo, el número de taxones es menor debido a que no se tomaron en cuenta todas aquellas URLs que contenian una extensión pdf, docx o ppt, ya que generalmente se encontraban encriptadas, para compensar esta situación se utilizó 7 veces más el número de URLs que en el de la prueba manual.

Por otra parte, sólo once patrones se utilizaron para extraer los taxones de la prueba supervisada o semi-automatizada. El número de taxones encontrados es del 45.56 % comparado con la prueba manual.

	Prueba manual	**Prueba supervisada**
Corpus usado	Primer corpus (corpus manual)	Tercer corpus (corpus automático)
Tiempo de elaboración	20 días	5 días
Número de taxones encontrados	79	36
Número de URLs visitadas	30	230
Número de niveles taxonómicos	8	3

Tabla 5.2: Comparación de resultados

Capítulo 6

Conclusiones

En este trabajo se conjuntaron la teoría de la extracción de CDs de [Alarcón, 2009] y relaciones léxicas de [Ortega, 2007] para el desarrollo de la herramienta de extracción de taxonomías utilizando corpus textuales con carácter científico de la web.

La extracción se realiza a partir de un término dado por el usuario con la opción de analizar corpus formales. Con ello, se cumplió el objetivo principal, los objetivos específicos y la motivación propuestos en esta tesis:

- Se diseñó una metodología supervisada para extraer taxonomías a partir de un término dado por el usuario de forma semi-automática en áreas de especialidad.
- Se aplicó la metodología propuesta al área del PLN para el idioma español.
- Se extrajo la taxonomía para el área del PLN.
- La taxonomía supervisada fue evaluada por los expertos en el área de PLN.
- Se desarrolló una herramienta que opera con el método supervisado para la extracción de taxonomías.
- Se extrajeron los términos asociados al término semilla.

Con los resultados obtenidos se comprobó que la metodología desarrollada para la extracción de taxonomías es una forma viable de obtenerlas, ya que los taxones extraídos fueron validados como satisfactorios. Ésto se debe que con un conjunto de patrones léxicos se pueden obtener los suficientes taxones para organizar la información. El uso de las relaciones léxicas es adecuado para ligar los términos y los

espacios entre las palabras. Además se observó que con pocos patrones léxicos se pueden encontrar taxones para el área del PLN.

Para encontrar un mayor cúmulo de términos jerarquizados se necesitarían más patrones léxicos. Esto para tener una mayor cobertura en los resultados obtenidos y por consiguiente, más términos ubicados en la taxonomía. Sin embargo, con unos cuantos patrones léxicos se pueden obtener taxones relacionados a una área del conocimiento.

La herramienta desarrollada permite extraer taxonomías de cualquier dominio del conocimiento (en español) a partir de un término semilla. También puede extraer taxonomías de cualquier idioma que use el espacio para separar las palabras, sólo cambiando los patrones léxicos a los del idioma requerido.

Al generar una taxonomía no es necesario hacer uso del etiquetado de las partes de la oración para obtener los taxones. Aunque para tener mayor precisión en los términos obtenidos puede hacerse uso de este método, claro que con ello aumentaría el tiempo de elaboración de la taxonomía.

Por otra parte, en las frases extraídas con el término semilla se encontró que es necesario saber cuales son los límites por la derecha y por la izquierda del término. El límite por la derecha del término se ve afectado, la mayoría de las veces, por el uso de «de», «para» y «¿qué..?» porque éstos influyen el sentido del término, por lo que pueden referirse a otro término o realizar una pregunta.

Aunque las taxonomías son estáticas, la información siempre está en constante incremento. Entonces para obtener una taxonomía actualizada es necesario tener un corpus elaborado a partir de la web para hacer un análisis semi-automático con la herramienta propuesta en esta tesis y así obtener la taxonomía del término semilla.

Una ventaja de la representación de las taxonomías es que permiten abarcar una gran cantidad de información en un menor espacio. Además brindan apoyo en el proceso de inidzación de textos y en la recuperación de información.

La taxonomía se puede usar como una herramienta de apoyo para la enseñanza y aprendizaje del algún dominio del conocimiento porque facilita la asimilación de la información. Además es un método para ordenar y organizar la información, también el algoritmo para la generación de taxonomías optimiza la recuperación de la información en la web o de algún corpus.

En la comparación entre los resultados manuales y semi-automáticos se observa una mejoría en el tiempo de extracción de taxonomías. En otras palabras, la elaboración de la taxonomía semi-automática llevo menos tiempo y menos esfuerzo que la

taxonomía manual, lo que facilitó el trabajo en la extracción de la misma.

Por útlimo, el elemento para afinar en el ECODE es la conversión de PDF a TXT, con el fin de mejorar sus resultados.

6.1. Aportaciones

Se obtuvo una taxonomía del PLN en español a partir de un corpus generado de la web con ayuda de la herramienta desarrollada en esta tesis, usando el enfoque de extracción de relaciones léxicas.

Además, esta herramienta es de ayuda en el ánalisis manual del comportamiento de los patrones léxicos en las frases que contienen el término semilla en los diferentes dominios del conocimiento, así como en la organización del conocimiento la cual podría emplearse como ayuda en el área de bibliotecología.

Por lo tanto, con los archivos generados por la herramienta desarrolla pueden ayudar a mejorar la extracción del cúmulo de términos asociados al término semilla y por consiguiente mejorar la extracción de taxonomías.

Con la metodología propuesta se pueden realizar búsqueda especializadas de un término. En otras palabras, se pueden encontrar áreas del conocimiento específicas y sus temas relacionados.

6.2. Trabajos futuros

Observar el comportamiento de las frases que contienen el término semilla apoyándose de la teoría de la RST [1] en español, para mejorar los resultados en la extracción de los hipónimos o hiperónimos y sinónimos de un término. También se podría mejorar la precisión de los marcadores metalingüísticos y lingüísticos.

Hacer una investigación del uso de los signos de puntuación para la relación de hiponímia, hiperonímia y sinonímia entre los términos. También estudiar la viabilidad de implementarlos en esté método para mejorar su funcionamiento.

Incrementar los pares de preposiciones que anteceden al término semilla e influyen en el sentido de la frase, empezando por «qué es el» y «para el» para desartar todos

[1] Mann, William C., and Sandra A. Thompson. Rhetorical Structure Theory: A Theory of Text Organization. Available as Information Sciences Institute Research Report 87-190, 4676 Admiralty Way, Marina del Rey, 82 pg. `http://www.sfu.ca/rst/01intro/intro.html`

aquellas frases que no definan a los términos.

Aplicar la programación recursiva en aquellas frases que contienen el término semilla y alguna de las relaciones léxicas sobre los taxones asociados al término semilla, para tener una mayor profundidad en cada taxonomía, es decir, para tener más de tres niveles taxonómicos.

Implementar en el extractor taxonómico un algoritmo de aprendizaje para obtener más patrones léxicos utilizando como fuente Ortega (2007) y así encontrar más taxones que conformen una taxonomía. Realizar una interfaz gráfica de esta herramienta para facilitar su uso. Por último, explorar el uso de taxonomías en la generación de resumenes automáticos.

Bibliografía

Alarcón, Rodrigo: "*ECODE : ExtractordeContextosDefinitorios.*" *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios.* Tesis doctoral, España, México: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Instituto de Ingeniería, Universidad Nacional Autónoma de México., 2009.

Alarcón, Rodrigo; Bach, Carme y Sierra, Gerardo Eugenio: «Extracción de contextos definitorios en corpus especializados: Hacia una elaboración de una herramienta de ayuda terminográfica». *Revista Española de Lingüística (RSEL)*, 2007, **37**, pp. 247–277.

Amini, Amineh; Wah, Teh Ying; Saybani, Mahmoud Reza y Yazdi, Saeed Reza Aghabozorgi Sahaf: «A Study of Density-Grid based Clustering. Algorithms on Data Streams». *Fuzzy Systems and Knowledge Discovery (FSKD), 2011 Eighth International Conference*, 2011, **3**, pp. 1652 – 1656.

Amsler, Robert Alfred: *The structure of the Merriam Webster Pocket Dictionary.* Tesis doctoral, Deparmente of Computer Sciences University of Texas at Austin, 1980.

> Notas: los pasos por Jhon Olney se encuentran en la página 35 del documento

Arano, Silvia: «La ontología: una zona de interacción entre la Lingüística y la Documentación.[en línea]. EN: Hipertext.net (2)». `http://www.hipertext.net` [Consulta: 20 septiembre 2012], 2003.

Arano, Silvia: «Los tesauros y las ontologías en la Biblioteconomía y la Documentación. [en línea]. EN: Hipertext.net (3)». `http://www.hipertext.net/web/pag260.htm` [Consulta: 18 septiembre 2012], 2005.

Argudo, Sílvia y Centelles, Miquel: «Metodología para el diseño de taxonomías corporativas». *Investigación Bibliotecológica*, 2005, **19(039)**, pp. 158–177.

Aude, Marie y Soto, Michel: *"Metadata- and Ontology-Based Semantic Web Mining". Web semantics ontology.* pp. 259–296. Idea Group Publishing, 2006.

Notas: Mapas RDF página 5

Auger, Alain y Barriére, Caroline: «"Pattern-based approaches to semantic relation extraction. A state-of-the-art". Terminology: international journal of theoretical and applied issues in specialized communication». *Terminology*, 2008, **14**, pp. 1–19.

Barceló, Miquel: *"La representación del conocomiento.Ïnteligencia artificial.* Editorial UOC, 2009.

Baroni, Marco y Bisi, Sabrina: «Using cooccurrence statistics and the web to discover synonyms in a technical language». En: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC).*, pp. 1725–1728. Lisbon, Portugal, 2004.

Berland, Matthew y Charniak, Eugene: «Finding parts in very large corpora.» En: *Proceedings of the 37th Annual Meeting of the Association for Computacional Linguistics.*, pp. 57–64, 1999.

Bilgin, Arif; Ellson, John; Gansner, Emden; Hu, Yifan; North, Stephen; Koren, Yehuda; Dobkin, David; Dwyer, Tim; Koutsofios, Eleftherios; Lilly, Bruce; Low, Glen; Mocenigo, John; Scheerder, Jeroen; Woodhull, Gordon y Caldwell, Don: «Graphviz-Graph Visualization Software», 1999. `http://www.graphviz.org/`.

Bird, Steven; Klein, Ewan y Loper, Edward: *"Language Processing and Python ". Natural Language Processing with Python.* O'Reilly Media, Inc., 2009.

Cabré, M. Teresa: *La terminología: Representación y comunicación.* IULA Universidad Pompeu Fabra. Segunda reimpresión, 1999.

Caraballo, Sharon A.: «Automatic construction of a hypernym-labeled noun hierarchy from text». *ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, 1999, pp. 120–126.

Castillo, Luis F.; Franco, Oscar H. y Giraldo, Jaime A.: «Agentes basados en Ontologías para la Web Semántica. Euro-American conference on telematics and information systems.», 2010.

Centelles, Miquel: «Taxonomías para la categorización y la organización de la información en sitios web. [en línea]. EN: Hipertext.net (3)». `http://www.hipertext.net.` [Consulta: 19 julio 2012], 2005. Consulta 19 julio 2012.

Cimiano, Philipp; Hotho, Andreas y Staab, Steffen: «Çomparing conceptual, divisive and agglomerative clustering for learning taxonomies from text"». En: *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*, pp. 435–439, 2004.

Correa, Gabriela: «Economías basadas en el conocimiento: Alberta y México». *Denarius: revista de economía y administración.*, 2006, **12**, pp. 221–243.

Cózar, Ramón López: *.Análisis y gestión del diálogo". Los sistemas de diálogo.* Universidad Autónoma de Barcelona, 2006.

Díaz, Flor Nancy; Joyanes, Luis y Medina, Víctor Hugo: «Taxonomía, ontología y folksonomía, ¿qué son y qué beneficios u oportunidades presentan para los usuarios de la web?» *Universidad y Empresa, Redalyc*, 2009, **8(16)**, pp. 242–261. Universidad del Rosario.

Desongles, Juan: *Çonceptos básicos sobre la organización de la información..Ayudantes Tecnicos de Informática de la Junta de Andalucia.* MAD-Eduforma, 2005.

documents, Solid: «Solid Converter». `http://www.perucomputec.com/yoel/solid-converter-portable/`, 2010.
`http://www.perucomputec.com/yoel/solid-converter-portable/`

Dolan, William; Vanderwende, Lucy y Richardson, Stephen D.: «Automatically deriving structured knowledge bases from on-line dictionaries». *In Proceedings of the First Conference of the Pacific Association for Computational Linguistics (Vancouver, Canada),*, 1993, pp. 5–14.

Fedor, Alicia: *"La teoría general de la terminología".Terminología: Teoría y práctica.* Equinoccio, 1995.

Notas: página 54

Fernández, Anisleiby: «Organización de los contenidos en los sitios Web: las taxonomías.» *ACIMED*, 2007. [en línea]. EN: Hipertext.net 15(5). `http://bvs.sld.cu/revistas/aci/vol15_05_07/aci12507.htm` [Consulta: 28 septiembre 2012].

FERRATER, JOSÉ: *Diccionario de filosofía. Tomo 1 A-K.* Sudamericana Buenos Aires, quinta edición, Buenos Aires, Argentina, 1969.

GIANCOLI, C. DOUGLAS: *Çalor".Fisica. Principios con aplicaciones. 6 eds.* Pearson Educación„ 2007.

GIBERT, KARINA: «"Técnicas híbridas de Inteligencia Artificial y Estadística para el descubrimiento de conocimiento y la minería de datos".Tendencias de la Minería de Datos en España». *Red Nacional de MiDA (Red Nacional de Minera de Datos y Aprendizaje) Digital3, Sevilla*, 2004, pp. 119–130. `http://www.lsi.us.es/redmidas/LibroMD.htm http://sid.usal.es/idocs/F8/FDO24054/Memoria_2006_2008_Laboratorio.txt`.

GIMENO, JAVIER: *"Sistemas de indización aplicados en bibliotecas: clasificaciones, tesauros y encabezamiento de materias."Tratado Básico de Biblioteconomía.* Editorial Complutense, 2004.

GIRJU, ROXANA: «Automatic Detection of Causal Relations for Question Answering». *in Proceedings of ACL Workshop on Multilingual Summarization and Question Answering*, 2003, **12**, pp. 76–83.

GÓMEZ, ANDRÉS y DE JESÚS, IGNACIO: *"Datos y algoritmos avanzados.Ïntroducción a la Computación.* Cengage Learning Editores, 2008.

GONZALO, JULIO y VERDEJO, M. FELISA: *Recuperación y extracción de información."Tecnologías del lenguaje.* Editorial UOC, 2003.

GRISHMAN, RALPH: *.ªpplications". The Handbook of Computational Linguistics and Natural Language Processing.* John Wiley & Sons, 2010.

HARPRING, PATRICIA: *Çontrolled Vocabularies in Context ", "What Are Controlled Vocabularies? ", Relationships in Controlled Vocabularies ".Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works.* Getty Publications, 2010. `http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/index.html`. `http://www.getty.edu/research/publications/electronic_publications/intro_controlled_vocab/what.html`

HEARST, MARTI A.: «Automatic acquisition of hyponyms from large text corpora». En: *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pp. 539–545. Nantes, France, 1992.

Hernández, Ariadna Carolina: *Análisis lingüístico de definiciones analíticas para la búsqueda de reglas que permitan su delimitación automática.* Tesis de Licenciatura, UNAM, México, 2009.

INTECO: *"La taxonomía 2.0". TAXONOMIA (v 2.0) DE SOLUCIONES DE SEGURIDAD TIC.* INTECO (Instituto Nacional de Tecnologías de la Comunicación), 2009.

Jackson, Peter y Moulinier, Isabelle: *"Document retrieval", Ïnformation extraction". Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization.* Natural Language Processing. John Benjamins Pub., 2007.

Lancaster, Frederick Wilfrid: *"Búsqueda con lenguaje natural y el vocabulario postcontrolado". El control del vocabulario en la recuperación de información (2a ed.).* Universitat de València, 2002.

Lin, Dekang; Zhao, Shaojun; Qin, Lijuan y Zhou, Ming: «Identifying synonyms among distributionally similar words». En: *Proceedings of the 18th international joint conference on Artificial intelligence*, pp. 1492–1493, 2003.

Lorente, Mercè: «Ontología sobre economía y recuperación de información.[en línea]. EN: Hipertext.net (3)». `http://www.hipertext.net.` [Consulta: 20 septiembre 2012], 2005.

Malaise, Véronique; Zweigenbaum, Pierre y Bachimont, Bruno: «"Detecting Semantic Relations between Terms in Definitions".» pp. 55–62. 3rd edition of CompuTerm Workshop (CompuTerm 2004) at Coling 2004., Geneva, Switzerland, 2004.

Martí, María Antonia: *Tecnologías del lenguaje.* Editorial UOC, 2003.

Meyer, Ingrid: *Recent Advances in Computational Terminology.* capítulo .Extracting knowledge-rich contexts for terminography. A conceptual and methodological framework", pp. 279–302. John Benjamins Publishing, 2001.

Miller, George A.: «What is WordNet? [en línea]». `http://wordnet.princeton.edu/` [Consulta: 10 octubre 2012], 1980.
`http://wordnet.princeton.edu/`

Miller, George A.; Beckwith, Richard; Fellbaum, Christiane; Gross, Derek y Miller, Katherine J.: «Introduction to WordNet: An On-line Lexical Database». *International Journal of Lexicography*, 1999, **3(4)**, pp. 235–312.

Miranda, Alice: *"Lenguajes alfabéticos de clasificación". Procesamiento de la Información en Bibliotecología.* EUNED. San José : Universidad Estatal a Distancia., 1995.

Miranda, Juan Carlos; Lobo, Roque Antonio; Castro, Dario Alberto; Mendoza, Anibal y Gracerant, Oswaldo: *Manual de laboratorio de física mecánica.* Uninorte, 2010.

Montiel, Mayra: *"Principios generales de taxonomía". Introducción a la Flora de Costa Rica.* Editorial Universidad de Costa Rica, 1991.

Montoya, Hugo Humberto: *"Taxonomía". Microbiología básica para el área de la salud y afines. Segunda edición.* Universidad de Antioquia, 2008.

Moreiro, José Antonio: *"La representación y recuperación de los contenidos digitales: de los tesauros conceptuales a las folksonomías". Tendencias en documentación digital.* pp. 81 – 109. Editorial UOC, 2006.

Notas: Capítulo 3

Moreiro, José Antonio; Sánchez, Sonia y Morato, Jorge: «Panorámica y tendencias en topic maps». [http://www.hipertext.net.[], 2003. http://www.upf.edu/hipertextnet/numero-1/topic_maps.html

Munoz, Marcia y Nagarajan, Ramya (Eds.): *Sentence Segmentation tool.* University of Illinois at Urban-Champing, Cognitive Computation Group, 2001. Segmentador oracional. http://cogcomp.cs.illinois.edu/page/tools_view/2

Noruzi, Alireza: «Folksonomies: Why do we need controlled vocabulary? [en línea]. EN: Webology 4(2)». http://www.webology.ir/2007/v4n2/editorial12.html [Consulta: 23 agosto 2012], 2007.

Oliver, Antoni; Moré, Joaquim y Climent, Salvador: *"Los corpus lingüísticos". Traducción y tecnologías.* Editorial UOC, 2008.

Ortega, Rosa María: *Descubrimiento Automático de Hipónimos a partir de Texto no Estructurado.* Tesina o Proyecto, Tesis de maestría,Puebla, México. Instituto Nacional de Astrofísica, Óptica y Electrónica, 2007.

Ortega, Rosa María; Aguilar, César Antonio; Villaseñor, Luis; Montes, Manuel y Sierra, Gerardo Eugenio: «Hacia la identificación de relaciones de hiponimia/hiperonimia en Internet». http://www.scielo.cl/scielo.php?pid=S0718-09342011000100006&script=sci_arttext [Consulta: 29 septiembre 2012], 2011.

Osorio, Fray León: *Introducción a la Programación en Java.*, 2007.

Notas: Ejemplos usados

Osorio, Fray León: *Lógica y programación orientada a los objetos: un inicio al desarrollo de software*, 2008.

Pantel, Patrick y Pennacchiotti, Marco: «Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations». En: *Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL)*, pp. 113–120. Sydney, Australia, 2006.

Pasca, Marius: «Finding Instance Names and Alternative Glosses on the Web: WordNet Reloaded». *In CICLing*, 2005, **LNCS 3406**, pp. 280–292. Berlin/Heidelberg: Springer-Verlag.

Peis, E.; Herrera-Viedma, E.; Hassan, Y. y Herrera, J. C.: «Análisis de la web semántica: estado actual y requisitos futuros». *El Profesional de la Información*, 2003, **12(5)**, pp. 368–376.

Pereira, Fernando; Tishby, Naftali y Lee, Lillian: «Distributional clustering of english words». *Proceedings of the 31st Annual Meeting of the Association for Computational Lingistics (ACL)*, 1993, pp. 183–190.

Pino, Raúl; Gómez, Alberto y De Abajo, Nicolás: *Ïnteligencia artificial". Introducción a la inteligencia artificial: Sistemas expertos, redes neuronales artificiales y computación evolutiva.* Servicio de Publicaciones, Universidad de Oviedo, 2001.

Quero, Adriana: *Definición de una ontología para la guía de conocimiento swebok.* Tesis de maestría, Mérida, Venezuela. Universidad de los Andes., 2007.

RAE: «Diccionario de la Lengua Española [versión 22 electrónica].» `http://www.rae.es`, 2011.

Ramírez, Sandra: «Linneo: la pasión de un médico por la clasificación de los seres vivos». *Ciencia de la Salud*, 2007, **5(1)**, pp. 515–517.

Ravichandran, Deepak y Hovy, Eduard: «Learning Surface Text Patterns for a Question Answering System». *1n Proceedings of ACL-2002*, 2002, pp. 41–47.

Riloff, Ellen y Shepherd, Jessica: «A Corpus-based approach for building semantic lexicons». En: *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 117–124, 1997.

Rodríguez, Keilyn y Ronda, Rodrigo: «Web semántica: un nuevo enfoque para la organización y la recuperación de información en el Web [en línea]». *Acimed http://www.hipertext.net. [Consulta: 8 noviembre 2012]*, 2005, **13**. `http://bvs.sld.cu/revistas/aci/vol13_6_05/aci030605.htm`

Rueger, Stefan: *Multimedia Information Retrieval*. Morgan & Claypool Publishers, 2010.

Salvat, Guiomar y Serrano, Vicente: *Cómo empezó todo: la emergencia y consolidación de la Sociedad de la Información ."La revolución digital y la Sociedad de la Información*. Comunicación Social, España, 2011.

Schwab, Didier; Lafourcade, Mathieu y Prince, Violaine: «Antonymy and conceptual vectors.» En: *Proceedings of Computational Linguistics*, pp. 904–910. Taipei, Taiwan, 2002.

Schwitter, Rolf: «Controlled Natural Languages». `https://sites.google.com/site/controllednaturallanguage/`, 1999.

Schwitter, Rolf: «Controlled Natural Languages for Knowledge Representation». *Proceedings of COLING 2010, Beijing, China*, 2010. Centre for Language Technology Macquarie University.

Sáez, Miguel: *Diseño de un sistema de extracción de información de artículos de Wikipedia*. Proyecto Fin de Carrera, España. Universidad Carlos III de Madrid. Departamento de Ingeniería Telemática, 2009. Universidad Carlos III de Madrid Escuela Politécnica Superior.

Shapiro, Stuart Charles y Eckroth, David: *Encyclopedia of artificial intelligence, Volume 1*. Wiley, 1992.

Sicard, Andrés y Vélez, Mario Elkin: «Universalidad de la computación cuántica geométrica: modelo de tres estados». *Ingeniería y Ciencia*, 2005, **1(1)**, p. 6.

Sierra, Gerardo Eugenio: «Diseño de corpus textuales para fines lingüísticos». En: *IX Encuentro Internacional de Lingüística en el Noroeste*, pp. 445–462, 2008.

Sierra, Gerardo Eugenio: «Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos». *Linguamática*, 2009, **2(1)**, pp. 13–37. Revista 1.

Sierra, Gerardo Eugenio; Alarcón, Rodrigo; Aguilar, César Antonio y Barrón, Alberto: *Towards the building of a corpus of definitional contexts.*

volumen 1. en Proc. XII EURALEX International Congress, Volume I, Turín, Italia, 2006.

Sinclair, John: «Preliminary recommendations on Corpus Typology. EAG–TCWG–CTYP/P.[en línea]». `http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html` [Consulta: 30 agosto 2012], 1996. Technical report EAGLES (Expert Advisory Group on Language Engineering Standards). `http://www.ilc.cnr.it/EAGLES/corpustyp/corpustyp.html`

Smith, Barry: «.ºntology in Information Systems". Blackwell Guide to the Philosophy of Computing and Information. [en línea]». `http://ontology.buffalo.edu/ontology%28PIC%29.pdf` [Consulta: 23 agosto 2012], 2001.

Soler, Concha y Gil, Isidoro: «Posibilidades y límites de los tesauros frente a otros sistemas de organización del conocimiento: folksonomías, taxonomías y ontologías». *Revista Interamericana de Bibliotecología*, 2010, **33**, pp. 361–377.

Soler, María Concepción: *Evaluación de vocabularios controlados en la indización de documentos mediante índices de consistencia entre indizadores.* Tesis doctoral, Universidad Politecnica de Valencia, Valencia, España, 2009.

Soriano, Edmundo Pavel: *Clasificación de opiniones mediante aprendizaje de máquinas: el caso de reseñas sobre películas.* Tesis de licenciatura, D.F. México. Universidad Nacional Autónoma de México, 2011.

Sowa, John F.: *Knowledge Representation. Logical, Philosophical and Computational Foundations.* Brooks Cole Publishing Co; Pacific Grove, CA, 2000.

Torruella, Joan y Llisterri, Joaquim: «Diseño de corpus textuales y orales». En: *Filología e informática. Nuevas tecnologías en los estudios filológicos*, pp. 45–77. Editorial Milenio y Universidad Autónoma de Barcelona, Barcelona, 1999.

Trimble, Louis: *English for Science and Technology: A Discourse Approach.* Cambridge University Press, Cambridge, 1985.

Turney, Peter: «Mining the Web for synonyms: PMI-IR versus LSA on TOEFL». En: *Proceedings of the 12th European Conference on Machine Learning.*, pp. Freiburg, Germany. 491–502. Freiburg, Germany, 2001.

(U.S.), National Information Standards Organization: *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.* NISO Press, 2005. ANSI/NISO Z39.19-2005.

Vieyra, José Luis: *Adaptación, optimización y expansión de ECODE, un sistema extractor de contextos defnitoiros.* Tesis de licenciatura, D.F. México. Universidad Nacional Autónoma de México, 2011.

Villayandre, Milka: *"Los corpus..Aproximación a la lingüística computacional.* Tesis doctoral, León. Universidad de León. Departamento de Filología Hispánica y Clásica, 2010.

Winston, Morton E.; Chaffin, Roger y Herrmann, Douglas: «A Taxonomy of Part-Whole Relations». *Cognitive Science*, 1987, **11**, pp. 417–444.

Yule, George: *"Semántica". El lenguaje. Tercera edición. Traducido porNuria Bel Rafecas.* Ediciones AKAL, Móstoles, Madrid, 2007.

Zesch, Torsten y Gurevych, Iryna: «Analysis of the Wikipedia Category Graph for NLP Applications». En: *Proceedings of the Second Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing. Association for Computational Linguistics*, pp. 1–8. Association for Computational Linguistics, Germany, 2007.