



SEP
SECRETARÍA DE
EDUCACIÓN PÚBLICA



cenidet[®]
*Centro Nacional de Investigación
y Desarrollo Tecnológico*

Centro Nacional de Investigación y Desarrollo Tecnológico

Departamento de Ciencias Computacionales

TESIS DE MAESTRÍA EN CIENCIAS

Aprendizaje ontológico a partir de contextos definitorios

Presentada por

Samuel Hipólito Rocha García

Ing. en Sistemas Computacionales por el I.T. de Minatitlán

Como requisito para la obtención del grado de:

Maestría en Ciencias en Ciencias de la Computación

Director de tesis:

Dr. Noé Alejandro Castro Sánchez

Co-Director de tesis:

Dra. Azucena Montes Rendón

Jurado:

Dr. Juan Gabriel González Serna

Dr. José Alejandro Reyes Ortiz



SEP
SECRETARÍA DE
EDUCACIÓN PÚBLICA



TECNOLÓGICO NACIONAL DE MÉXICO

Centro Nacional de Investigación y Desarrollo Tecnológico

"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cuernavaca, Morelos a 06 de agosto del 2019
OFICIO No. DCC/076/2019

Asunto: Aceptación de documento de tesis

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO
PRESENTE

Por este conducto, los integrantes de Comité Tutorial del Ing. Samuel Hipólito Rocha García, con número de control M16CE021, de la Maestría en Ciencias de la Computación, le informamos que hemos revisado el trabajo de tesis profesional titulado "Aprendizaje ontológico a partir de patrones definitorios" y hemos encontrado que se han realizado todas las correcciones y observaciones que se le indicaron, por lo que hemos acordado aceptar el documento de tesis y le solicitamos la autorización de impresión definitiva.

DIRECTOR DE TESIS

Dr. Noé Alejandro Castro Sánchez
Doctor en Ciencias de la
Computación
08701806

CO-DIRECTORA DE TESIS

Dra. Azucena Montes Rendón
Doctora en Ciencias
4001014

REVISOR 1

Dr. José Alejandro Reyes Ortiz
Doctor en Ciencias de la
Computación
8457365

REVISOR 2

Dr. Juan Gabriel González Serna
Doctor en Ciencias de la
Computación
7820329

C.p. M.E. Guadalupe Garrido Rivera - Jefa del Departamento de Servicios Escolares.
Estudiante
Expediente

NACS/lmz



"2019, Año del Caudillo del Sur, Emiliano Zapata"

Cuernavaca, Mor.,
No. de Oficio:
Asunto:

5/septiembre/2019
SAC/260/2019
Autorización de
impresión de Tesis

ING. SAMUEL HIPÓLITO ROCHA GARCÍA
CANDIDATO AL GRADO DE MAESTRO EN CIENCIAS
DE LA COMPUTACIÓN
PRESENTE

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado "Aprendizaje ontológico a partir de patrones definitorios", ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE

Excelencia en Educación Tecnológica®
"Conocimiento y tecnología al servicio de México"

DR. GERARDO VICENTE GUERRERO RAMÍREZ
SUBDIRECTOR ACADÉMICO



SEP TecNM
CENTRO NACIONAL
DE INVESTIGACIÓN
Y DESARROLLO
TECNOLÓGICO
SUBDIRECCIÓN
ACADÉMICA

C.p. M.E. Guadalupe Garrido Rivera.- Jefa del Departamento de Servicios Escolares.
Expediente

GVGR/mcr

Dedicatoria

Dedico este trabajo de tesis a mis padres quienes me han apoyado siempre y me han dado el conocimiento de las cosas más importantes que se puedan aprender: el cariño, el respeto, los valores y el amor por el estudio.

También a mis hermanos que siempre me apoyan y confían en mí, sé que puedo contar con ellos y con su particular forma de ser lo cual los hace únicos y valiosos.

Agradecimientos

Quiero expresar mi más profundo agradecimiento a la Dra. Azucena Montes Rendón. Su dedicación al conocimiento científico y su espíritu enérgico y afectivo son un ejemplo a seguir y han incitado mi ilusión por el estudio.

Al director de esta tesis, el Dr. Noe Castro Alejandro Sánchez, quién ha sido un importante guía en cada paso de esta investigación, no sólo con sus conocimientos sino también con su apoyo en diferentes situaciones y su interés por promover la investigación.

A mis revisores de tesis: el Dr. José Alejandro Reyes Ortiz y el Dr. Juan Gabriel González Serna, quienes dedicaron parte de su tiempo a realizar las revisiones necesarias para poder realizar un trabajo digno de una investigación de maestría. Gracias por su ayuda apoyo y consejos.

A mis compañeros que los considero como una auténtica fortuna repartida en varios sitios distintos. Muchas gracias a todos mis amigos, a los de siempre que han sobrellevado con gran paciencia las innumerables horas de mi ausencia y a los de ahora, que espero conservar por mucho tiempo, ambos grupos forman parte inseparable en esta etapa de mi vida.

Al Centro Nacional de Investigación y Desarrollo Tecnológico, y al Tecnológico Nacional de México por aceptarme como alumno, que gracias a sus profesores tengo las herramientas necesarias para poder ser un investigador de éxito y gracias al apoyo económico brindado por el CONACYT pude dedicarme íntegramente al estudio de mi carrera.

Índice

Dedicatoria	5
Agradecimientos.....	6
Índice de tablas	9
Índice de figuras	9
Resumen	11
Capítulo 1. Introducción.....	13
1.1. Descripción del problema	14
1.2. Antecedentes.....	15
1.2.1 Creación Automática de Ontologías a partir de Textos	15
1.2.2. Poblado automático de ontologías espaciales a partir de texto no estructurado	16
1.3. Objetivo.....	17
1.3.1. Objetivo general	17
1.3.2. Objetivos específicos	17
1.4. Alcances y limitaciones.....	17
1.4.1. Alcances.....	17
1.4.2. Limitaciones	17
Capítulo 2. Estado del arte	18
2.1. Definitional verbal patterns for semantic relation extraction	18
2.2. Developing a definitional knowledge extraction system	18
2.3. Researching specialized languages.....	20
2.4. The role of verbal predications.....	21
2.5. Pattern construction for extracting domain terminology	23
2.6. A lexico-semantic pattern language for learning ontology instances from text	24
2.7. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos	25
2.8. Análisis, diseño e implementación de un agente deliberativo para extraer contextos definitorios en textos especializados	26
2.9. Comparación de las técnicas	27

Capítulo 3. Marco teórico	29
3.1. Extracción y recuperación de información	29
3.2. Definición	29
3.3. Contextos definitorios	30
3.4. Patrones definitorios	31
3.4.1. Patrones tipográficos.....	31
3.4.2. Patrones sintácticos	32
3.5. Relaciones semánticas	32
3.6. Procesamiento del Lenguaje Natural.....	33
3.7. Freeling.....	34
3.8. Autómata finito.....	35
3.9. Ontología	36
3.9.1. Aprendizaje ontológico	37
Capítulo 4. Método de solución	38
4.1. Módulo de identificación de patrones en el texto	38
4.1.1. Autómata para la identificación de contextos definitorios	39
4.1.2. Autómata para la identificación del término y género próximo	41
4.2. Módulo para determinar la relación de hiperonimia	42
4.3. Ejemplo del funcionamiento del autómata	43
4.4. Plataforma web.....	47
Capítulo 5. Pruebas.....	49
5.1. Etiquetado manual	49
5.2. Identificación de contextos definitorios	50
5.3. Identificación del término y género próximo.....	52
5.4. Resultados de las pruebas	54
5.4.1. Identificación de contextos definitorios.....	54
5.4.2. Identificación del término y género próximo	55
5.5. Limitaciones detectadas	55
5.5.1. Identificación de contextos definitorios.....	56
5.5.2. Identificación de término y género próximo.....	57
Capítulo 6. Conclusiones y trabajos futuros	58

Referencias	59
Anexos	61
afasia.txt	61
afta.txt	62
alergia.txt	63

Índice de tablas

Tabla 1. Ejemplo de identificación de eventos [1]	15
Tabla 2. Evaluación de la extracción de PVD [7].....	19
Tabla 3. Evaluación de la identificación de los CD [7]	20
Tabla 4. Ejemplos de predicaciones verbales [9]	22
Tabla 5. Estructuras de las predicaciones verbales	22
Tabla 6. Resultados de la evaluación [10].....	24
Tabla 7. Comparación del estado del arte.....	28
Tabla 8. Símbolos aceptados por el autómata principal.....	40
Tabla 9. Símbolos aceptados por el autómata secundario.....	41
Tabla 10. Categorías gramaticales.....	43
Tabla 11. Archivo generado por la aplicación.....	48
Tabla 12. Precisión y cobertura para la identificación de contextos definitorios	54
Tabla 13. Eficacia en la identificación del término y género próximo	55

Índice de figuras

Figura 1. Cognición de eventos [1].....	15
Figura 2. Arquitectura de ECODE [3]	18
Figura 3. Diagrama de la metodología para extraer terminología de un dominio [10]	23
.....	
Figura 4. Diagrama de la metodología para el aprendizaje de instancias de ontologías a partir de textos [11]	25
Figura 5. Diagrama de procesamiento con GATE [16].....	27
Figura 6. Simbología en la comparación del estado del arte.....	28
Figura 7. Elementos de una definición [3]	29
Figura 8. Elementos principales de un CD [20]	31
Figura 9. Metodología de solución	38
Figura 10. Autómata para la identificación de contextos definitorios.....	39
Figura 11. Autómata para para la identificación del término y género próximo	41

Figura 12. Relación de hiperonimia	43
Figura 13. Flujo del autómata, parte 1	44
Figura 14. Flujo del autómata, parte 2	44
Figura 15. Flujo del autómata, parte 3	45
Figura 16. Flujo del autómata, parte 4	46
Figura 17. Flujo del autómata, parte 5	46
Figura 18. Pantalla principal	47
Figura 19. Pantalla de resultado	47
Figura 20. Opciones al pie del resultado	48
Figura 21. Elementos principales de un CD [20]	49
Figura 22. Etiquetado manual para la identificación de contextos definitorios .	51
Figura 23. Resultado para la identificación de contextos definitorios	52
Figura 24. Etiquetado manual para la identificación del término y género próximo	53
Figura 25. Resultado para la identificación del término y género próximo	53

Abstract

Ontology learning is the study of methods for the automatization of ontology creation and population. The information in a discourse domain can be organized and structured through an ontology, which exposes its most relevant terms and the relationships that exist between them.

In turn, definitional contexts are pieces of text that contain a definition, a clear and precise description of the meaning of a word. Finding definitional contexts in a text will not only identify relevant terms in a topic; these text fragments offer a taxonomic structure themselves, which will facilitate the ontological learning tasks.

The purpose of this work is to expose a method that identifies definitional contexts present in unstructured text files, and then carries out the automatic extraction of the existing semantic relationships in the identified text fragments, which can be used to form taxonomies.

The process is carried out by means of a deterministic finite automaton designed from predefined patterns that describe the structure of the defining contexts that present an analytical definition.

The system developed was evaluated by means of precision and recall tests, which showed a value of 90.24% in accuracy and 74% in coverage. As a result of this method, we have a tool to assist in the structuring of texts, which can mainly benefit automatic ontology population.

Resumen

El aprendizaje ontológico es el estudio de los métodos para la creación y poblado automático de ontologías. La información de un dominio de discurso puede ser organizada y estructurada mediante una ontología, la cual expone sus términos más relevantes y las relaciones que existen entre ellos.

A su vez, los contextos definitorios son fragmentos de texto que contienen una definición, una descripción clara y precisa del significado de una palabra. Encontrar los contextos definitorios en un texto no sólo permitirá identificar términos relevantes en un tema; estos fragmentos de texto ofrecen una estructura taxonómica en sí mismos, lo cual facilitará las tareas de aprendizaje ontológico.

El propósito de este trabajo es exponer un método que identifique los contextos definitorios presentes en archivos de texto no estructurado, para después llevar a cabo la extracción automática de las relaciones semánticas existentes en los fragmentos de texto identificados, las cuales se pueden utilizar para formar taxonomías.

El proceso se lleva a cabo por medio de un autómata finito determinista diseñado a partir de patrones predefinidos que describen la estructura de los contextos definitorios que presentan una definición analítica.

El sistema desarrollado fue evaluado mediante pruebas de precisión y cobertura, las cuales demostraron un índice de 90.24% en precisión y 74% en cobertura. Como resultado de este método se tiene una herramienta para asistir a la estructuración de textos, la cual podrá beneficiar principalmente al poblado automático de ontologías.

Capítulo 1. Introducción

Desde la creación del internet, el acceso a la información se ha vuelto cada vez más sencillo, y la cantidad de información disponible ha crecido de manera exponencial. Procesar tanta información de manera manual eventualmente se volvería problemático, lo que ha dado paso al surgimiento de la web semántica.

La idea de la web semántica es añadir información semántica adicional al contenido existente en la *world wide web* para que esta pueda ser fácilmente interpretada por programas de computadora, lo que a su vez permitirá un procesamiento más rápido de la información y búsquedas más precisas.

Se han desarrollado diversas herramientas para describir los recursos de la web de manera semántica, entre ellas las ontologías, que permiten describir los conceptos de un dominio de estudio en particular y las relaciones que existen entre estos. El aprendizaje ontológico consiste en la creación automática de dichas ontologías, lo cual por lo general se logra haciendo uso de técnicas de procesamiento del lenguaje natural.

Dado que las ontologías frecuentemente describen relaciones jerárquicas entre los conceptos de un dominio, uno de los recursos de mayor utilidad para esta tarea son las definiciones, y más concretamente, los contextos definatorios.

En el presente documento se explicará una propuesta del proceso necesario para la búsqueda e identificación de contextos definatorios en textos de especialidad, la identificación de la relación que existen entre los conceptos clave de estos contextos, así como la creación de taxonomías a partir de estas relaciones.

En el marco teórico se hablará más a detalle de los principales recursos que se manejan en este trabajo, como las ontologías, el procesamiento del lenguaje natural y los contextos definatorios.

1.1. Descripción del problema

Las ontologías se han convertido en un recurso importante en los últimos años, ya sea para la descripción de los conceptos de un dominio, así como para añadir información semántica adicional a los recursos de la web, permitiéndonos acercarnos aún más a la visión de la web semántica.

Lamentablemente, el proceso de creación y poblado de ontologías de manera manual es complicado, requiere de tiempo y esfuerzo, y requiere de expertos en el dominio de la ontología a realizar. Es por ello que se recurre al aprendizaje ontológico, el cual permite la creación automática de ontologías haciendo uso de técnicas de procesamiento del lenguaje natural. Sin embargo, existen pocas técnicas de aprendizaje ontológico para la clasificación de información en idioma español, y no son muy eficientes o están incompletas.

Una forma de facilitar las tareas de aprendizaje ontológico es a partir de taxonomías generadas en base a relaciones de hiperonimia, donde una palabra general deriva en términos más específicos. Este tipo de relaciones pueden ser encontradas en las definiciones de tipo analítica. En este documento se propone un método para la identificación semiautomática de contextos definitorios que contienen una definición analítica.

1.2. Antecedentes

A continuación, se detallan los trabajos previos sobre aprendizaje ontológico realizados en el Cenidet que han servido como fundamento para esta investigación.

1.2.1 Creación Automática de Ontologías a partir de Textos

En una investigación previa hecha en [1], se utilizaron técnicas del procesamiento del lenguaje natural para el análisis y estructuración de textos en ontologías, con base en un sistema de eventos.



Figura 1. Cognición de eventos [1]

En términos simples, el trabajo define que en un texto pueden ocurrir múltiples eventos, los cuales se relacionan con un actor, un tiempo, un lugar y una causa. En general, un evento está identificado por un verbo, por lo que puede haber tantos eventos como verbos haya en un texto [1].

Al identificar un evento en una oración, se procede a identificar al actor de dicho evento, el lugar y el momento en que sucede el evento, si es que existen en el fragmento de texto que se está analizando. En la tabla 1 se muestra un ejemplo de la identificación de eventos en tripletas.

Tabla 1. Ejemplo de identificación de eventos [1]

Texto de entrada	Israel lanza bombardeos contra objetivos de Hizbulá
Eventos y actantes identificados	Israel lanza bombardeos contra objetivos de Hizbulá

Tripletas obtenidas	<p>lanza instancia_de EventoVerbal Israel instancia_de Agente bombardeos instancia_de Objeto objetivos de Hizbulá instancia_de Beneficiario lanza realizado_por Israel lanza tiene_objeto bombardeos lanza beneficia_a objetivos de Hizbulá</p>
---------------------	---

El enfoque utilizado hace uso de patrones semánticos para la extracción de eventos verbales y nominales, con un índice de precisión de 97% para la detección de eventos verbales y de 42% para eventos nominales, debido al problema de la polisemia en los sustantivos.

Los eventos verbales son aquellos que están asociados a una acción representada por un verbo. Los eventos nominales entonces son aquellos que están asociados a una acción representada por un sustantivo.

1.2.2. Poblado automático de ontologías espaciales a partir de texto no estructurado

En esta investigación se describe la metodología para el poblado de la ontología “OntoEvento” mediante la identificación de entidades espaciales en un texto no estructurado [2].

El proceso que sigue consta de cuatro pasos y hace uso de algunas herramientas existentes para la identificación de patrones sintácticos.

1. Etiquetado de eventos. Se toma una noticia de la web en texto plano y se realiza el etiquetado de eventos.
2. Reconocimiento de entidades espaciales mediante Calais. La noticia etiquetada en el paso anterior se ingresa a la herramienta Calais y, con el apoyo de una base de datos con los nombres de países y principales ciudades de todo el mundo, se realiza la detección de entidades espaciales.
3. Reconocimiento de entidades espaciales mediante patrones. Se complementa el paso anterior haciendo uso de patrones sintácticos para la detección de las entidades espaciales que no fueron detectadas, y se definen los pares Evento – Espacio.
4. Creación de tripletas. Se obtienen las tripletas “evento *tiene_espacio* espacio” para el poblado de la ontología espacial pasando las entidades espaciales a sus correspondientes clases.

Las pruebas para el reconocimiento de entidades espaciales demostraron un índice de precisión y cobertura de 94%, mientras que las pruebas del poblado ontológico demostraron un índice de precisión y cobertura de 80%.

1.3. Objetivo

1.3.1. Objetivo general

Diseñar un método que realice aprendizaje ontológico a partir de textos no estructurados, en idioma español, a través de la identificación y uso de contextos definitorios.

1.3.2. Objetivos específicos

- Obtener un banco de información de texto no estructurado, en idioma español.
- Formalizar los patrones definitorios.
- Identificar los patrones en el texto no estructurado.
- Crear relaciones entre los términos de un dominio.

1.4. Alcances y limitaciones

1.4.1. Alcances

- Extraer la información desde una base de conocimientos.
- Determinar la relación de hiperonimia entre los elementos identificados.

1.4.2. Limitaciones

- El banco de información será del dominio de medicina.
- El texto analizado será en idioma español.

Capítulo 2. Estado del arte

En este capítulo se describirán las investigaciones recientes más relevantes relacionadas con el aprendizaje ontológico y la extracción de relaciones semánticas en un texto.

2.1. Definitional verbal patterns for semantic relation extraction

Objetivo: extraer relaciones semánticas a partir de textos en español, mediante el uso de patrones verbales definitorios [3]–[6].

Qué se hizo: se realizó una extensa investigación sobre los componentes de los Contextos Definitorios y las principales relaciones semánticas encontradas en éstos. Se clasificaron los Patrones Verbales Definitorios correspondientes a los diferentes tipos de definiciones. Se desarrolló un sistema para la búsqueda automática y reconocimiento de Patrones Verbales Definitorios, filtrado de contextos no relevantes, y la identificación de elementos constitutivos: término, definición y PVD. Se puede observar la arquitectura del sistema desarrollado en la Figura 2.

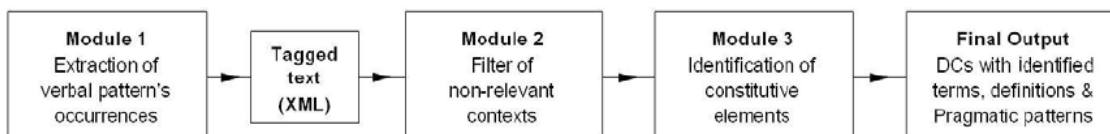


Figura 2. Arquitectura de ECODE [3]

Resultados: se describe el rol de los patrones verbales definitorios para extraer conocimiento definitorio. Se describe también un set de relaciones semánticas que conectan a una definición con patrones verbales específicos en un contexto definitorio. Con esta información, se diseñó un sistema automático para la extracción de conocimiento definitorio, el cual será de ayuda para la extracción de relaciones semánticas en español.

2.2. Developing a definitional knowledge extraction system

Objetivo: desarrollar un sistema para la extracción automática de conocimiento definitorio a partir de corpus especializado [4], [7].

Qué se hizo: se describe a los Contextos Definitorios (CD) y su estructura, la cual está compuesta por un Término y una Definición, generalmente delimitados por elementos tipográficos o Patrones Verbales Definitorios (PVD). En esta investigación se seleccionaron 15 PVDs para su extracción en un corpus técnico de

documentos especializados en Leyes, Genoma, Economía, Medio ambiente, Medicina, Informática, y Lenguaje general.

Los PVD fueron delimitados entre las etiquetas “<dvp></dvp>”, y el texto antes y después de ellos fueron delimitados entre las etiquetas “<left></left>” y “<right></right>” respectivamente.

Después, se realizó un filtrado de los Contextos No Definitivos, al detectar la ocurrencia de ciertas partículas gramáticas o secuencias sintácticas donde un PVD no es usado para definir un término.

Finalmente se identificaron los elementos de los CD mediante un árbol de decisiones. Este es capaz de detectar, mediante inferencias lógicas, la posición probable de los términos, definiciones y patrones pragmáticos dentro del CD.

Resultados: la evaluación de la metodología se llevó a cabo en dos partes. Primero se evaluó la extracción de PVD y filtrado de contextos no relevantes, mediante las medidas de Precisión y Cobertura. Los resultados se pueden ver en la Tabla 2.

Tabla 2. Evaluación de la extracción de PVD [7]

Verbal pattern		P	R
Concebir (como)	To conceive (as)	0.67	0.98
Definir (como)	To define (as)	0.84	0.99
Entender (como)	To understand (as)	0.34	0.94
Identificar (como)	To identify (as)	0.31	0.9
Consistir de	To consist of	0.62	1
Consistir en	To consist in	0.6	1
Constar de	To comprise	0.94	0.99
Denominar también	Also denominated	1	0.87
Llamar también	Also called	0.9	1
Servir para	To serve for	0.55	1
Significar	To signify	0.29	0.98
Usar como	To use as	0.41	0.95
Usar para	To use for	0.67	1
Utilizar como	To utilise as	0.45	0.92
Utilizar para	To utilise for	0.53	1

Después se evaluó la identificación de los elementos de un CD, asignando un valor dependiendo de la calidad de la clasificación, donde:

- 3 – Contextos donde los elementos constitutivos fueron clasificados correctamente.
- 2 – Contextos donde los elementos constitutivos fueron clasificados correctamente, pero también se clasificó información adicional.

1 – Contextos donde los elementos constitutivos no fueron clasificados correctamente.

∅ – Contextos que el sistema no pudo clasificar.

Los resultados se pueden observar en la Tabla 3.

Tabla 3. Evaluación de la identificación de los CD [7]

Verbal pattern	3	2	1	∅
Concebir (como)	68.57	15.7	11.4	4.28
Definir (como)	65.1	18.2	10.4	6.25
Entender (como)	54.16	20.8	8.33	16.66
Identificar (como)	51.72	5.17	34.5	8.62
Consistir de	60	0	20	20
Consistir en	60.81	8.1	15.5	15.54
Constar de	58.29	23	2.97	15.74
Denominar también	21.42	28.6	7.14	42.85
Llamar también	30	40	0	30
Servir para	53.78	27.3	0.01	18.18
Significar	41.26	44.4	3.17	11.11
Usar como	63.41	14.6	17.1	4.87
Usar para	36.26	33	4.39	26.37
Utilizar como	55.1	28.6	10.2	6.12
Utilizar para	51.51	19.7	10.6	18.18

Los resultados se consideran generalmente satisfactorios, pero aún queda mucho trabajo por hacer para mejorar el desempeño de las inferencias del árbol de decisiones.

2.3. Researching specialized languages

Objetivo: Crear un repositorio de Patrones Léxico Sintácticos asociados a las estructuras ontológicas que describen, para facilitar el desarrollo de ontologías [8].

Qué se hizo: Existen ciertas unidades lingüísticas que proveen información importante acerca de diferentes tipos de relaciones entre las palabras. En esta investigación se estudian los Patrones Léxico Sintácticos (PLS), los cuales proveen información acerca de las estructuras conceptuales presentes en las ontologías, también llamados Patrones de Diseño Ontológico (PDO).

Identificar relaciones conceptuales por medio de patrones lingüísticos

Las relaciones conceptuales representan la unión entre dos o más unidades de conocimiento en un campo de estudio. Los patrones lingüísticos que describen

estas relaciones conceptuales pueden ser aplicados para la extracción de terminología.

Patrones léxico sintácticos en Ingeniería del Conocimiento

El desarrollo de ontologías requiere el descubrimiento de conceptos de un dominio, sus propiedades y cómo se relacionan con otros conceptos. Para facilitar el desarrollo de ontologías, se han aplicado patrones lingüísticos para la extracción de elementos ontológicos. Existen diversos patrones verbales y no-verbales para identificar relaciones de hiponimia, meronimia, agencia, causa, entre otros. El enfoque de esta investigación es el uso de patrones lingüísticos con el propósito de ayudar a los usuarios en la creación de ontologías

Estrategias para la extracción de PLS

1. Seleccionar patrones verbales en la literatura y adaptarlos a un esquema de notación basado en la Forma Backus-Naur.
2. Identificar los conceptos ontológicos, y buscar en la Web los constructos verbales que los vinculen de forma onomasiológica.
3. Buscar en documentos enciclopédicos los constructos verbales que vinculen los conceptos de acuerdo a relaciones ontológicas.

Resultados: se desarrolló un repositorio de PLS y PDO donde se reúnen algunas de las formas más usuales en las que un lenguaje puede expresar las estructuras ontológicas consideradas soluciones de modelado consensual. Aunque a lista de PLS no es exhaustiva, intenta representar las formas más comunes en las que un lenguaje puede expresar las estructuras identificadas.

2.4. The role of verbal predications

Objetivo: desarrollar un sistema automático de extracción de contextos definitorios, usando un set de reglas y restricciones obtenidas mediante el análisis de patrones recurrentes [9].

Qué se hizo: Los contextos definitorios están compuestos por secuencias llamadas *patrones*, como son: **T** (término), **D** (definición), **tm** (marcas tipográficas), **VP** (predicación verbal) y **PP** (predicación pragmática).

El enfoque principal de esta investigación son las Predicaciones Verbales, las cuales se clasifican en dos tipos:

- La forma simple, que hace uso de un verbo y ocasionalmente una partícula gramática.
- La forma compleja, que hace uso del pronombre “**se**” junto con un verbo.

Algunos ejemplos de predicaciones verbales se pueden observar en la Tabla 4.

Tabla 4. Ejemplos de predicaciones verbales [9]

Formas simples	Formas complejas
Afirma que	Se basa en
Comprende	Se concibe como
Consiste en	Se conoce (como/con)
Consta de	Se considera (como)
Constituye	Se define como
Corresponde a	Se denomina (como)
Define a	Se encarga de
Incluye	Se refiere a
Ocurre	Se utiliza (para/en)

Haciendo una búsqueda selectiva en el Corpus de Referencia del Español Actual, se identificaron las distintas estructuras de las predicaciones verbales correspondientes a 10 verbos utilizados frecuentemente en contextos definitorios. Estas estructuras se muestran en la Tabla 5.

Tabla 5. Estructuras de las predicaciones verbales

Grupos		
se le denomina se denomina + art indef. se denomina + art def.	se conoce como se le* conoce se conoce* con	consiste en consiste básicamente en
se define como se define por se puede definir se define + art def. se le define se debe definir se ha logrado definirlos	se refiere* a se refiere* + art def.	Permite
se entiende + art def. se entiende como se entiende por se entiende cuando	comprende + art comprende desde	representa a incluye a

Resultados: se determinó qué predicaciones verbales ofrecen mejores resultados en la búsqueda de contextos definitorios. Para ello se realizó un etiquetado manual y se comparó contra un etiquetado automático, luego se calcularon las medidas de precisión y cobertura. La medida de cobertura para estas predicaciones va desde 50% hasta 100%, mientras que la precisión obtenida llegó a tener valores desde 10% hasta 100%. Aún se necesita estudiar las combinaciones necesarias de los elementos constitutivos, y considerar otros elementos característicos, como las marcas tipográficas y las predicaciones pragmáticas.

2.5. Pattern construction for extracting domain terminology

Objetivo: desarrollar una metodología para la obtención automática de patrones (Patrones básicos y Patrones Verbales Definitorios) para extraer la terminología del dominio [10].

Qué se hizo: se seleccionaron 120 documentos en español con temas correspondientes a la Base Curricular y Propiedad del Plan de Estudio “D” de la carrera de Informática de la Universidad Agraria de La Habana, enfocándose en aquellos textos con información representativa, revisados y aprobados por expertos de su dominio.

La metodología propuesta se compone de cuatro pasos:

1. Selección del corpus perteneciente al dominio
2. Anotación semi-automática de los términos pertenecientes al dominio en cuestión, con el apoyo de la herramienta TermEt y validación de un experto humano.
3. Obtención de los patrones básicos, extrayendo las etiquetas obtenidas de un análisis morfológico a las palabras anotadas.
4. Obtención de los Patrones Verbales Definitorios.

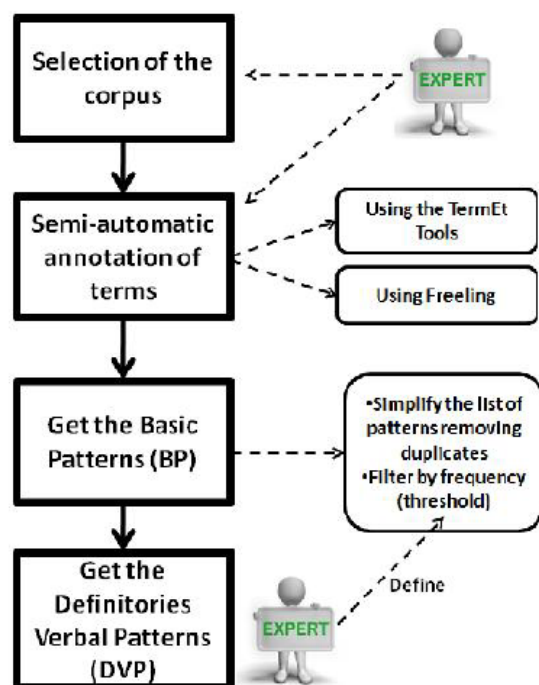


Figura 3. Diagrama de la metodología para extraer terminología de un dominio [10]

Resultados: la metodología obtenida fue evaluada en el dominio de ciencias de la computación y obtuvo un 97% en el caso de los patrones básicos y un 98% en el

caso de los patrones verbales definitorios. Después se evaluó la metodología en otros tres dominios con resultados similares, como se observa en la tabla 2.

Tabla 6. Resultados de la evaluación [10]

Dominio	Patrones básicos	PVD
Ciencias de la computación	97%	98%
Ingeniería agrícola	96%	97%
Medicina veterinaria	98%	98%
Agronomía	96%	96%

2.6. A lexico-semantic pattern language for learning ontology instances from text

Objetivo: desarrollar un método basado en reglas para aprender instancias de ontologías a partir de un texto [11].

Qué se hizo: el Lenguaje de Extracción de Información Hermes (HIEL) emplea conceptos semánticos de una ontología. El lenguaje es evaluado en el contexto de extracción de eventos y relaciones en noticias, como extensión al framework de personalización de noticias Hermes existente.

Hermes es un framework que puede ser utilizado para construir un servicio de noticias personalizado, a partir de la selección de conceptos de una base de conocimientos.

Con base en el lenguaje HIEL, se implementó el Motor de Extracción de Información Hermes (HIEE).

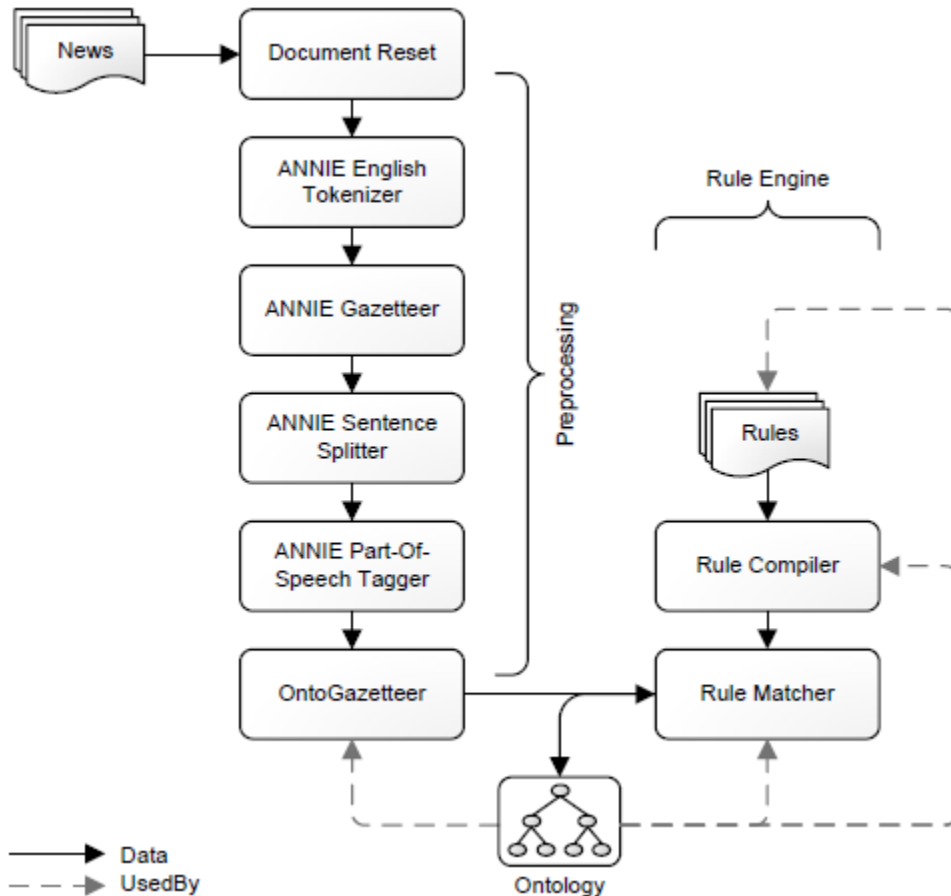


Figura 4. Diagrama de la metodología para el aprendizaje de instancias de ontologías a partir de textos [11]

Resultados: se evaluó la implementación del método propuesto construyendo reglas y midiendo el rendimiento de la extracción de eventos y relaciones usando estas reglas. En dos grupos de datos separados y ontologías correspondientes de los dominios financieros y políticos, esto resultó en una precisión de aproximadamente 80% y cobertura del 70%, ya que los patrones léxico-semánticos son superiores a los patrones léxico-sintácticos con respecto a expresividad.

2.7. Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos

Objetivo: presentar una síntesis de la investigación realizada por el Grupo de Ingeniería Lingüística del Instituto de Ingeniería en la UNAM, referente a la extracción automática de contextos definitorios en textos de especialidad en español, mediante el reconocimiento y análisis de patrones lingüísticos [12]–[15].

Qué se hizo: la investigación abarca diversos campos, empezando con la Extracción de Información y más concretamente la Extracción de Información

Terminológica y Conceptual (EITC), un campo orientado a la elaboración de ontologías y diccionarios electrónicos.

El estudio de Alarcón analiza los trabajos existentes en el ámbito de la extracción de Contextos Definitivos (CD), que son fragmentos textuales que aportan información para comprender el significado de un término. Se explica la estructura de los CD y su clasificación, así como las distintas clases de patrones que existen para su identificación y delimitación en un texto.

Finalmente, el ECODE desarrollado por Alarcón extrae los CD y los clasifica según el tipo de definición.

Para llevar a cabo los diversos estudios realizados, se conformó el corpus de experimentación, de prueba y de evaluación, tratando de ser consistentes en el empleo de las mismas fuentes.

Resultados: se desarrollaron diversas aplicaciones que hicieran uso de la metodología desarrollada para extraer CD de textos de especialidad a partir de patrones verbales. Las tres principales desarrolladas en el Grupo de Ingeniería Lingüística son las siguientes.

- Bancos de conocimiento: las Bases de Conocimiento Léxico son sistemas de bases de datos que almacenan, administran y proporcionan conocimientos obtenidos del lenguaje natural, a partir de textos tales como diccionarios, glosarios, artículo, etc.
- Extracción de relaciones léxicas: las relaciones léxicas son un tipo de relación producida a partir del significado que contiene una palabra, y pueden servir para representar el sistema de conceptos de un campo de conocimiento específico.
- El sistema Describe: es una aplicación cliente-servidor orientada a web que permite la clasificación y agrupamiento de definiciones encontradas en internet.

2.8. Análisis, diseño e implementación de un agente deliberativo para extraer contextos definitivos en textos especializados

Objetivo: mostrar el análisis, diseño e implementación de un agente deliberativo, con un mecanismo de aprendizaje supervisado, que permite identificar contextos definitivos en textos especializados [16].

Qué se hizo: se utilizó la herramienta GATE (General Architecture for Text Engineering) y el sistema ANNIE (A Nearly-New Information Extraction System) para extraer los contextos definitivos de manera automática.

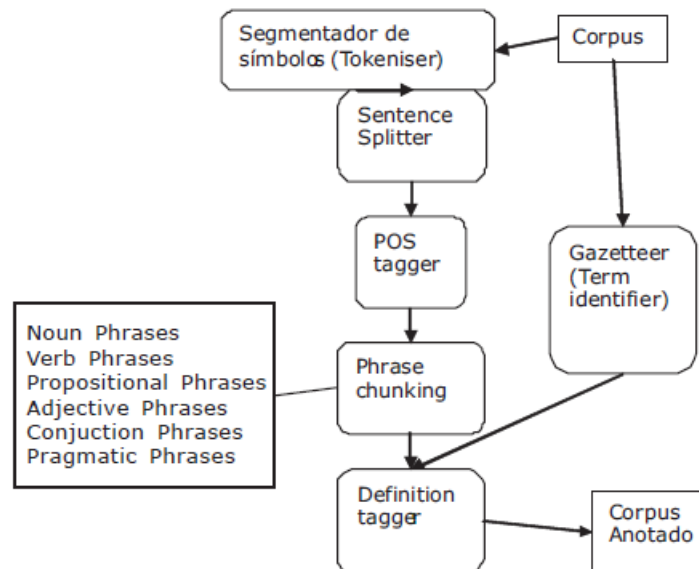


Figura 5. Diagrama de procesamiento con GATE [16]

Haciendo uso de las diferentes funcionalidades de las herramientas, se identificaron los patrones verbales y luego los contextos definitorios. Después, se utilizó la metodología GATE para la construcción del agente deliberativo.






El agente deliberativo interactúa con el corpus que ha sido analizado previamente con la herramienta GATE y, como resultado de dicha interacción, extrae los contextos definitorios de manera semiautomática.

Resultados: se realizó la descripción de una serie de patrones lingüísticos a partir de un corpus del dominio de las enfermedades neurológicas, lo cual permitió llegar a una representación formal para la extracción de CD. Igualmente se logró implementar el agente buscador sobre la plataforma JADE, de tal manera que arrojará como resultado archivos que pueden servir como corpus iniciales, para generar la extracción de definiciones del sistema global.

2.9. Comparación de las técnicas

En la Tabla 7 se describen las características que aborda cada uno de los trabajos investigados, así como las que se abordan en esta investigación. En la Figura 6 se describe el significado de los símbolos utilizados en la Tabla 7 (símbolos cortesía de <https://www.flaticon.com>).

Tabla 7. Comparación del estado del arte

							
Definitional verbal patterns for semantic relation extraction [3].	✓		✓		✓		
Developing a definitional knowledge extraction system [7].		✓		✓			
Researching specialized languages [8].	✓						
The role of verbal predications [5].			✓	✓	✓		
Pattern construction for extracting domain terminology [10].		✓	✓			✓	
A lexico-semantic pattern language for learning ontology instances from text [11].							✓
Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos [12].				✓		✓	✓
Análisis, diseño e implementación de un agente deliberativo para extraer contextos definitorios en textos especializados [16].			✓	✓			
Aprendizaje ontológico a partir de contextos definitorios.	✓		✓	✓	✓		✓








	Extracción de relaciones semánticas
	Extracción de PVD
	Uso de PVD
	Identificación de CD
	Identificación de elementos constitutivos
	Extracción de terminología del dominio
	Aprendizaje ontológico

Figura 6. Simbología en la comparación del estado del arte

Capítulo 3. Marco teórico

En este capítulo se abordarán los temas y áreas de investigación más importantes para el propósito de esta investigación.

3.1. Extracción y recuperación de información

La extracción y recuperación de información consiste en técnicas utilizadas para hacer una lectura y análisis automático de textos u otras fuentes de información mediante una computadora. Por lo general se hace uso de técnicas de PLN (Procesamiento del Lenguaje Natural) para en análisis de los textos [17].

El objetivo general de la extracción y recuperación de información consiste en hacer un análisis previo de la información mediante un programa de computadora. Un propósito más específico sería lograr hacer inferencias en base al contenido lógico de los datos de entrada, consiguiendo así datos semánticamente bien definidos con respecto a categoría y contexto, información estructurada.

La extracción y recuperación de información es frecuentemente utilizada para el análisis de las enormes cantidades de información que existe en la *world wide web*.

3.2. Definición

Una definición es un fragmento de texto cuyo propósito es “describir con claridad, exactitud y precisión el significado de una palabra o la naturaleza de una persona o cosa” [18]. Una definición describe de manera exclusiva y precisa la comprensión de un concepto o término. En el contexto de la lexicografía computacional, Ann Copestake plantea que una definición es la descripción lingüística de un concepto asociado a una palabra [19] (en el caso de la terminología, esta palabra equivale a un término). Una definición ofrece información importante acerca de una cosa determinada, por lo que se han diseñado diversas herramientas para identificar definiciones automáticamente a partir de un texto.

Las definiciones se pueden clasificar en cuatro tipos dependiendo de los elementos que la componen [3].

Estas pueden contener un **género próximo**, es decir, una palabra similar al término que se va a definir; y una **diferencia específica**, la descripción que diferencia el término del Género próximo.

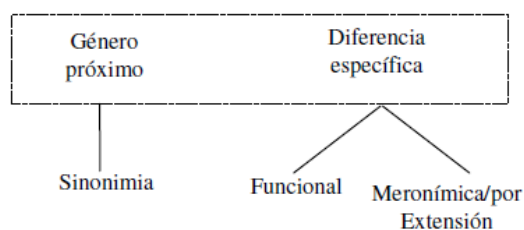


Figura 7. Elementos de una definición [3]

Los cuatro tipos de definiciones propuestas en [3] se presentan a continuación, seguidos de un ejemplo de cada tipo. El texto en color azul representa el *término* que se define, el texto en color rojo representa el *género próximo*, y el texto subrayado es la *diferencia específica*.

1. Definición analítica o aristotélica.

Ofrece explícitamente un género próximo y una diferencia específica.

Ejemplo:

Una mesa es un mueble compuesto por una tabla sostenida por una o más patas.

2. Definición sinonímica.

No ofrece ninguna diferencia específica, sino únicamente un género próximo que representa una unidad léxica que conceptualmente tiene una equivalencia con el término definido.

Ejemplo:

Una danza se le llama también a un baile.

3. Definición funcional.

Por medio de la diferencia específica ofrece una definición de la función, utilidad o fin de lo referido por el término.

Ejemplo:

Una tarjeta de débito se utiliza para hacer compras o disposiciones de dinero.

4. Definición extensional.

Por medio de la diferencia específica, y sin contar con género próximo, enumera las partes o componentes que forman un objeto.

Ejemplo:

Una computadora se compone de un gabinete, un monitor, un teclado y un ratón.

3.3. Contextos definatorios

Un contexto definatorio es un fragmento de texto que contiene información relevante para entender el significado de un concepto [20]. Según el trabajo sobre contextos definatorios hecho por Alarcón y Sierra, se sabe que:

- Las definiciones analíticas están organizadas con base en verbos que operan como núcleos.
- La predicación es una estructura del lenguaje natural ligada a las definiciones.
- La predicación es una estructura relacionada con la formulación y expresión de conceptos.

Los contextos definitorios poseen una estructura ya definida por [10], que se compone de un **Término**, una **Predicación verbal** o **Marcador tipográfico**, una **Definición** y **Patrones pragmáticos**, como se muestra en la Figura 8.

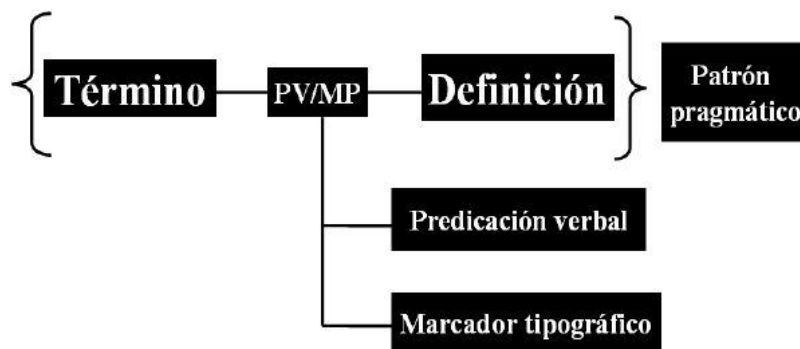


Figura 8. Elementos principales de un CD [20]

Los contextos definitorios se definen en función de dos elementos principales: un Término y una Definición, los cuales se encuentran conectados por un Patrón Definitorio (ya sea un patrón verbal o un marcador tipográfico), y en ocasiones pueden incluir un Patrón Pragmático que describe el contexto en el cual se define un término.

Adicionalmente, en [3] se consideraron dos tipos de predicaciones:

- a) Predicación simple o primaria: se encuentra conformada por un sujeto a la izquierda del verbo, y un predicado a la derecha del verbo.
- b) Predicación doble o secundaria: contiene un sujeto previo al verbo, y un objeto y predicado después del verbo.

3.4. Patrones definitorios

Los patrones definitorios son estructuras que conectan un término con su definición. Estos permiten la identificación de los tipos de definiciones encontradas en un texto, como se describe en [3]. Entre los principales tipos de patrones definitorios existen los patrones tipográficos y los patrones sintácticos.

3.4.1. Patrones tipográficos

Los patrones tipográficos son los indicadores visuales utilizados para resaltar y diferenciar el término que se va a definir del resto de la definición.

En el ejemplo:

***Gravedad.** Fuerza que sobre todos los cuerpos ejerce la Tierra hacia su centro.*

La palabra *gravedad* se encuentra señalada con el estilo tipográfico “negrita”.

3.4.2. Patrones sintácticos

Los patrones sintácticos son estructuras textuales encontradas con frecuencia como conexión entre el término y la definición.

Un tipo de patrón sintáctico es el **patrón verbal definitorio** (PVD), llamado así debido a que su componente principal es un verbo utilizado para abrir paso a una definición, como es el caso de los verbos *ser*, *definir* o *representar*. A estas palabras se les denomina **verbos definitorios**.

En el ejemplo:

*Una ecuación **es una** igualdad matemática entre dos expresiones [...]*

La estructura “es una” es un patrón verbal definitorio, cuyo elemento principal es el verbo definitorio “ser”.

Otro tipo de patrón sintáctico son los **marcadores reformulativos**, otra estructura textual utilizada para abordar el tema del discurso desde una perspectiva diferente. Algunos ejemplos de marcadores reformulativos son: *esto es*, *es decir*, *en otras palabras*, *por ejemplo*.

En el ejemplo:

*[...]los niveles de lengua se interrelacionan en el uso de una determinada lengua; **es decir**, que el análisis en niveles es únicamente metodológico.*

La estructura “es decir” es un marcador reformulativo que complementa la explicación previa sobre “los niveles de lengua”.

3.5. Relaciones semánticas

Las palabras pueden tener un significado dependiendo de si se ven de manera aislada o si se aprecia la manera en que se relacionan con otras palabras en un texto. El significado semántico de las palabras depende del contexto, y la relación que existe entre las palabras de una oración dotándolas de un significado específico, son las relaciones semánticas.

Las relaciones semánticas más comunes son:

- Meronimia / holonimia
- Hiponimia / hiperonimia
- Sinonimia / antonimia
- Paronimia

En [3] se identificaron tres tipos de relaciones semánticas principales con base en los componentes de *género próximo* y *diferencia específica* encontrados frecuentemente en distintos tipos de definiciones:

- **Hipónimo-Hiperónimo:** un hipónimo es una palabra que deriva de un término superior. Por ejemplo, **autobiografía** es hipónimo de **libro**.
- **Sinónimo:** los sinónimos son palabras que mantienen cierta equivalencia a nivel cognitivo. Por ejemplo, **danza** es sinónimo de **baile**.
- **Individuación:** se refiere a la obtención de entidades que pertenecen a un todo. Existen dos tipos de individuación:
 - a) De *cantidad/masa*, es decir, una relación entre una porción o pieza y una cierta sustancia o entidad. Por ejemplo, una **hora** es una porción de **tiempo**.
 - b) De *miembro/grupo*, es decir, una relación entre una entidad y el grupo al que pertenece. Por ejemplo, un **policía** es un miembro de la **fuerza policial**.

3.6. Procesamiento del Lenguaje Natural

El Procesamiento del Lenguaje Natural es un área de estudio relacionada con la interacción entre humanos y computadoras. Su objetivo principal es el diseño de programas que permitan a las computadoras entender e interpretar el lenguaje humano [3].

El aprendizaje automático es frecuentemente utilizado en los algoritmos de PLN, lo cual le permite a un programa tomar decisiones con base en estadísticas. Esto tiene la ventaja de que se puede aplicar fácilmente a sistemas grandes y complejos pues no hace falta especificar manualmente una larga lista de condiciones a evaluar.

El PLN tiene sus fundamentos en la lingüística, el estudio científico de las lenguas naturales y aspectos que se relacionan con ellas; su objetivo es diseñar una teoría de la estructura general de las lenguas naturales. La lingüística pretende describir las lenguas caracterizando el conocimiento que tiene los hablantes de ellas y determinar cómo las adquieren.

Entre las aplicaciones más comunes de PLN está la inteligencia artificial, el reconocimiento del habla, síntesis de voz, traducción automática y la extracción y recuperación de la información.

Existen diversas dificultades encontradas frecuentemente en el PLN, como la separación entre palabras, la recepción imperfecta de datos y la ambigüedad. Algunos de los problemas de ambigüedad son:

- Polisemia

El problema se presenta cuando una misma palabra tiene distintos significados.

- Anáfora

Se presenta cuando se hace referencia a un término anterior en un discurso.

- Catáfora

Cuando una palabra hace alusión a algo que está por venir en un discurso.

3.7. Freeling

Una de las herramientas más importantes para el desarrollo de esta investigación fue Freeling, una herramienta de análisis del lenguaje con la capacidad de realizar múltiples tareas de PLN en diferentes idiomas. Esta herramienta es desarrollada y mantenida por *TALP Research Center* en la Universidad Politécnica de Cataluña, donde múltiples personas han contribuido al proyecto mediante el desarrollo y mejoramiento de sus distintos módulos, así como creando y expandiendo sus bases de datos lingüísticas [21]–[25].

Los principales servicios que ofrece esta herramienta son:

- Tokenización de texto
- Separación de oraciones
- Análisis morfológico
- Tratamiento de sufijo, retoquenzación de pronombres clíticos
- Reconocimiento de palabras compuestas
- Reconocimiento flexible de términos multipalabra
- Separación de contracción
- Predicción probabilística de categorías de palabras desconocidas
- Codificación fonética
- Búsqueda basada en SED de palabras similares en un diccionario
- Detección de entidades nombradas
- Reconocimiento de fechas, números, ratios, moneda y magnitudes físicas
- Etiquetado PoS
- Análisis superficial basado en gráficos
- Clasificación de entidades nombradas
- Anotación de sentido y desambiguación basados en WordNet
- Análisis de dependencia basado en reglas
- Análisis de dependencia estadística
- Etiquetado semántico estadístico de roles
- Resolución de correferencia
- Extracción del gráfico semántico

En particular, el *etiquetado PoS* (Part of Speech), o simplemente *etiquetado gramatical*, fue de especial utilidad para esta investigación, ya que segmenta un texto por palabras y asigna una etiqueta a cada una en función de su categoría gramatical. Por ejemplo, para el siguiente fragmento de texto:

La alergia es una reacción inmunitaria del organismo frente a una sustancia generalmente inocua para el anfitrión.

Freeling realiza un análisis y genera un archivo con el etiquetado PoS como el siguiente:

```
La el DA0FS0
alergia alergia NCFS000
es ser VSIP3S0
una uno DI0FS0
reacción reacción NCFS000
inmunitaria inmunitario AQ0FS00
de de SP
el el DA0MS0
organismo organismo NCMS000
frente frente RG
a a SP
una uno DI0FS0
sustancia sustancia NCFS000
generalmente generalmente RG
inocua inocuo AQ0FS00
para para SP
el el DA0MS0
anfitrión anfitrión NCMS000
```

Donde cada línea del archivo contiene: La palabra analizada, su lema y su etiqueta PoS, basada en la propuesta de EAGLES [26]. Las etiquetas EAGLES PoS son de longitud variable y cada carácter corresponde a una característica morfológica. El primer carácter en la etiqueta es siempre la categoría (PoS). La categoría determina la longitud de la etiqueta y la interpretación de cada carácter en la etiqueta.

El etiquetado PoS hace uso no sólo de la definición de cada palabra, sino del contexto en el que aparecen, ofreciendo así un gran nivel de precisión para el etiquetado.

3.8. Autómata finito

Un autómata finito (AF) o máquina de estado finito es un modelo computacional que realiza cálculos en forma automática sobre una entrada para producir una salida.

Este modelo está conformado por 5 elementos:

- **Alfabeto.** Conjunto finito de símbolos aceptados por el autómata que formarán palabras o cadenas.
- **Conjunto de estados finito.** Los posibles estados en los que se puede encontrar un autómata.

- **Función de transición.** Describe el comportamiento del autómata dependiendo del símbolo leído y del estado en que se encuentra.
- **Estado inicial.** Estado en que el autómata se encuentra inicialmente.
- **Conjunto de estados finales.** Estados que provocan la parada del autómata.

El funcionamiento del autómata finito se basa en la función de transición, que recibe a partir de un estado inicial una cadena de caracteres pertenecientes al alfabeto (la entrada), y que va leyendo dicha cadena a medida que el autómata se desplaza de un estado a otro, para finalmente detenerse en un estado final o de aceptación, que representa la salida.

Formalmente, un **autómata finito** es una 5-tupla $(Q, \Sigma, q_0, \delta, F)$ donde:

- Q es un conjunto finito de estados;
- Σ es un alfabeto finito;
- $q_0 \in Q$ es el estado inicial;
- $\delta: Q \times \Sigma \rightarrow Q$ es una función de transición;
- $F \subseteq Q$ es un conjunto de estados finales o de aceptación.

Los autómatas se pueden clasificar en dos tipos:

- **Autómata Finito Determinista.** Autómata en el cual para cada estado en que se encuentre el autómata, y con cualquier símbolo del alfabeto leído, existe siempre a lo más una transición posible.
- **Autómata Finito No Determinista.** Autómata que posee al menos un estado tal que para un símbolo del alfabeto, existe más de una transición posible.

3.9. Ontología

Una ontología es la definición formal de tipos, propiedades y relaciones entre las entidades que existen en el dominio de un discurso particular [27].

Son ampliamente utilizadas en diversos campos, como la inteligencia artificial, web semántica, ingeniería de software y arquitectura de la información, con el propósito de reducir la complejidad y organizar la información.

Los componentes de una ontología son:

- **Individuos:** instancias u objetos.
- **Clases:** grupos o conceptos, similar a una clase en programación.
- **Atributos:** aspectos, propiedades o características de los objetos.
- **Relaciones:** formas en que las clases e individuos se relacionan.

- **Funciones:** estructuras complejas formadas a partir de ciertas relaciones que pueden ser utilizadas en lugar de un individuo.
- **Restricciones:** descripciones formales de algo que debe ser verdadero para que una afirmación sea aceptada como entrada.
- **Reglas:** oraciones en la forma de *si-entonces* que describen las inferencias lógicas que se pueden hacer a partir de ciertas afirmaciones.
- **Axiomas:** afirmaciones que comprenden la teoría descrita por la ontología en su campo de aplicación.
- **Eventos:** el cambio de atributos o relaciones.

3.9.1. Aprendizaje ontológico

El aprendizaje ontológico consiste en la creación automática o semiautomática de ontologías. Para lograr esto, se deben extraer términos del dominio correspondiente e identificar las relaciones entre estos conceptos a partir de un texto en lenguaje natural, y codificarlos con un lenguaje de ontologías para su futura recuperación [28].

El proceso generalmente se divide en ocho tareas principales, mas no todas se llevan a cabo en todos los sistemas de aprendizaje ontológico.

1. Extracción de la terminología del dominio
2. Descubrimiento de conceptos
3. Derivación de la jerarquía de conceptos
4. Aprendizaje de relaciones no-taxonómicas
5. Descubrimiento de reglas
6. Poblado de ontologías
7. Extensión de la jerarquía de conceptos
8. Detección de frame y eventos

Capítulo 4. Método de solución

El método diseñado para el propósito de esta investigación se compone de 2 módulos:

- Módulo para la identificación de patrones en el texto
- Módulo para determinar la relación de hiperonimia

El *módulo para la identificación de patrones en el texto* recibe como entrada un documento de texto no estructurado, identifica en él los contextos definitorios y sus componentes, y da como resultado un documento de texto procesado.

El *módulo para determinar la relación de hiperonimia* recibe como entrada el texto estructurado y proporciona como salida el documento de texto etiquetado en el cual se señalan las relaciones de hiperonimia.

El resultado de este método es un documento con relaciones taxonómicas bien definidas a partir de las cuales se pueden crear ontologías más complejas que describan los términos relevantes del texto original. En la Figura 9 se describe de manera visual el flujo de la metodología de solución.

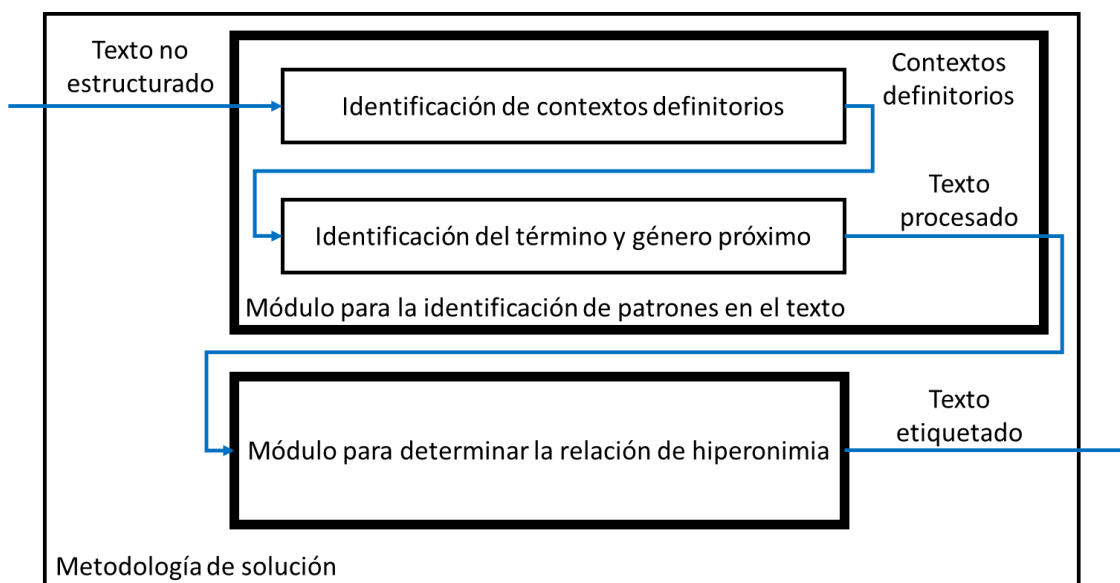


Figura 9. Metodología de solución

4.1. Módulo de identificación de patrones en el texto

Para la identificación de los contextos definitorios, un script en lenguaje Ruby procesa el texto mediante dos operaciones principales: el etiquetado POS del texto

con la herramienta Freeling, y el análisis del texto a través de un autómata finito determinista.

El etiquetado POS asigna una etiqueta a cada palabra y símbolo en el texto que representa su función dentro de una oración. Adicionalmente, Freeling separa el texto por oraciones. A continuación, el autómata analiza cada etiqueta como entrada, y para cada una produce un cambio de estado y una salida.

Dependiendo de las salidas producidas por el autómata, el script identifica los componentes de la definición para presentarlos al final de la ejecución. A su vez, el autómata se compone de dos partes: el autómata principal, el encargado de identificar los contextos definitorios válidos; y el autómata para identificar el término y género próximo, el cual entra en función durante la ejecución del autómata principal.

El autómata desarrollado consta de un conjunto de estados finito, el cual contiene un estado inicial, un estado final, y un conjunto de *estados de control*. Los estados de control son aquellos en los que el sistema realiza tareas específicas, las cuales se detallan en los siguientes párrafos.

4.1.1. Autómata para la identificación de contextos definitorios

El autómata principal sólo reconoce un fragmento de texto como un contexto definitorio si sus etiquetas conducen al estado final, el estado 18. En la Figura 10 se presenta un diagrama del autómata, y en la tabla 4 se describen las etiquetas POS utilizadas.

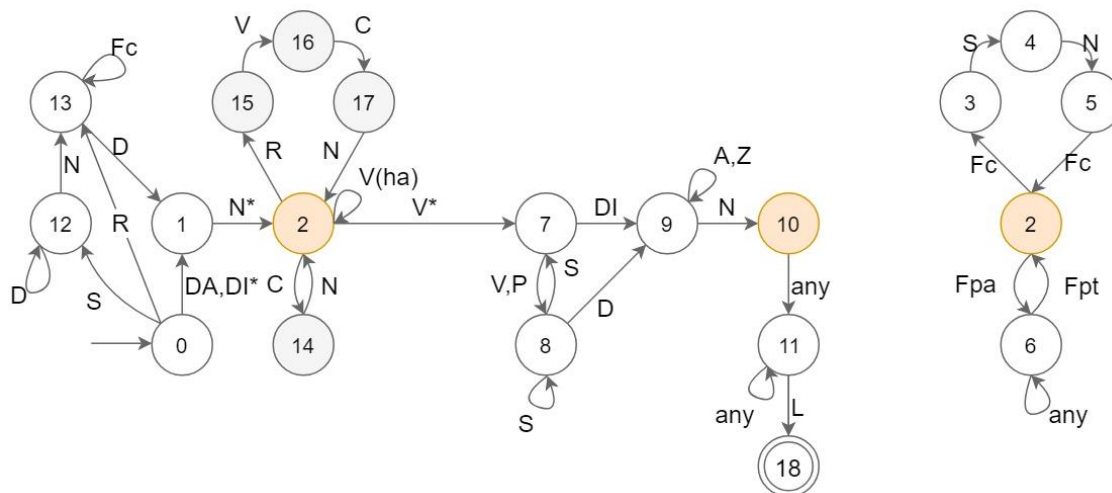


Figura 10. Autómata para la identificación de contextos definitorios

Tabla 8. Símbolos aceptados por el autómata principal

Etiqueta	POS
C	Conjunción
D	Determinante
Fc	Coma
Fpa	Paréntesis de apertura
Fpt	Paréntesis de cierre
L	Salto de línea
N	Sustantivo
P	Pronombre
R	Adverbio
S	Preposición
V	Verbo
<i>any</i>	Cualquier etiqueta

Especificaciones adicionales: este autómata cuenta con dos estados de control indicados en color amarillo, los estados 2 y 10. Cada vez que una transición conduce a uno de estos estados, se detiene el autómata principal y entra en función el autómata para identificar el término y género próximo, el cual se describe en el siguiente subtema.

También es importante mencionar que el sistema sólo acepta contextos definitorios que contienen una definición analítica, pues sólo estas presentan una relación de hiperonimia entre el género próximo y el término que se define. A su vez, se ha decidido utilizar sólo el verbo definitorio “ser” para estudiar el significado de este verbo en diferentes contextos. Se ha considerado además que el verbo “ser” está frecuentemente asociado con definiciones analíticas [29].

Transiciones específicas: adicionalmente, algunas transiciones tienen condiciones específicas. Estas transiciones se encuentran marcadas en el diagrama con un asterisco, y se explican a continuación.

Transición: $\delta(0, DA/DI)=1$

Condición: sólo si el lema no es “otro”.

Explicación: no es usual encontrar este término al inicio de un contexto definitorio, sólo causa la detección de múltiples falsos positivos.

Transición: $\delta(1, N)=2$

Condición: sólo si el lema no es “causa”.

Explicación: no es usual encontrar este término en esta posición en un contexto definitorio, sólo causa la detección de múltiples falsos positivos.

Transición: $\delta(2, V)=7$

Condición: sólo si el verbo no está en forma de gerundio.

Explicación: se detectó que las estructuras que presentan un verbo en gerundio no hacen referencia a un contexto definitorio válido.

4.1.2. Autómata para la identificación del término y género próximo

El siguiente autómata sólo reconoce un fragmento de texto como un término o género próximo válido si sus etiquetas conducen a uno de los estados de control (marcados en el diagrama en color verde). En la Figura 11 se presenta un diagrama del autómata, y en la tabla 5 se describen las etiquetas POS utilizadas.

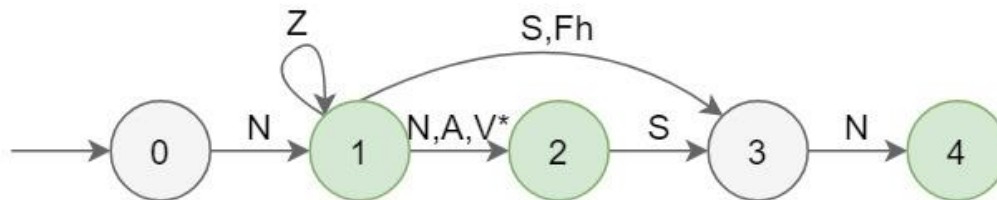


Figura 11. Autómata para para la identificación del término y género próximo

Tabla 9. Símbolos aceptados por el autómata secundario

Etiqueta	POS
A	Adjetivo
Fh	Diagonal (/)
N	Sustantivo
S	Preposición
V	Verbo
Z	Número

Especificaciones adicionales: el autómata inicia con la última etiqueta analizada por el autómata principal y finaliza en el momento en que la siguiente etiqueta no conduce a otro estado. Si en ese momento el autómata se encuentra en un estado de control, el autómata principal continúa normalmente desde el estado en el que se detuvo, utilizando la última etiqueta analizada por el autómata del término y género próximo. Este autómata cuenta con 3 estados de control indicados en color verde, los estados 1, 2 y 4.

Si la siguiente etiqueta no conduce a otro estado y el autómata no se encuentra en un estado de control, existen dos opciones.

- Utilizar el último término o género próximo válido.

- Descartar el contexto definitorio.

Ya que el sistema guarda el término o género próximo identificado cada que este autómata entra en un estado de control, es posible utilizar el último término identificado en caso de “error”. Por otra parte, si ningún término o género próximo ha sido identificado, se trata de un verdadero caso de error, por lo que el contexto definitorio es descartado y se devuelve el flujo del programa al autómata principal, el cual vuelve al estado inicial antes de continuar el análisis con la última etiqueta analizada por este autómata.

Transiciones específicas: algunas transiciones, marcadas con un asterisco, tienen condiciones específicas y se explican a continuación.

Transición: $\delta(1, V)=2$

Condición: sólo si el verbo no es “ser”.

Explicación: Freeing etiqueta algunos adjetivos como verbos, usualmente cuando son homónimos de un verbo.

4.2. Módulo para determinar la relación de hiperonimia

A partir de los contextos definitorios identificados, el sistema registra los términos en el orden que fueron encontrados. Como se mencionó anteriormente, para los contextos definitorios que contienen una definición analítica, el género próximo tiene una relación de hiperonimia con el término definido. En el caso de las definiciones analíticas siempre se presenta el término antes que el género próximo, por lo cual se puede determinar que el segundo término identificado siempre es hiperónimo del primero.

Por ejemplo, en la frase:

La hepatitis es una enfermedad inflamatoria que afecta a el hígado.

Se ha identificado el término “hepatitis” y el género próximo “enfermedad inflamatoria”. Podemos decir que la hepatitis es un tipo de enfermedad inflamatoria, es decir, se presenta una relación en la cual el género próximo es hiperónimo del término (Figura 12).

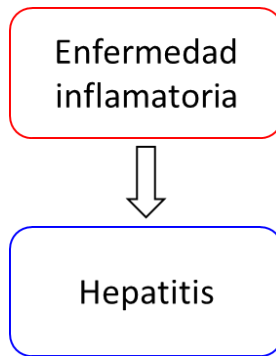


Figura 12. Relación de hiperonimia

Al final del análisis, el sistema presenta los elementos identificados: los contextos definitorios y su respectivo par de términos hipónimo-hiperónimo, como se describe en el capítulo 4.4. Plataforma web.

4.3. Ejemplo del funcionamiento del autómata

En las siguientes figuras se describe el flujo del autómata conforme se recibe la categoría gramatical de cada palabra como entrada. En este ejemplo se analiza el contexto definitorio que describe el significado de “exantema”. En color azul está señalado el término y en color rojo el género próximo. El progreso del flujo del autómata se indica mediante el subrayado. Debajo de cada palabra se indica su categoría gramatical con una letra mayúscula verde, según la Tabla 10.

Tabla 10. Categorías gramaticales

Letra	Categoría gramatical
D	Determinante
N	Sustantivo
V	Verbo
A	Adjetivo
P	Pronombre
C	Conjunción
S	Preposición
L	Punto y seguido

En la Figura 13 se muestra la primera etapa del flujo del autómata. Iniciando en el estado 0, se recibe “D” como entrada y se avanza al estado 1. Después se recibe “N” como entrada y se avanza al estado 2, el cual es un estado de control en el que el sistema transfiere el flujo al autómata para identificar el término y género próximo.

Un *exantema* es una *erupción cutánea* generalizada que aparece como
 D N V D N A V P V C
 N A S D N A
 manifestación clínica de una enfermedad sistémica.

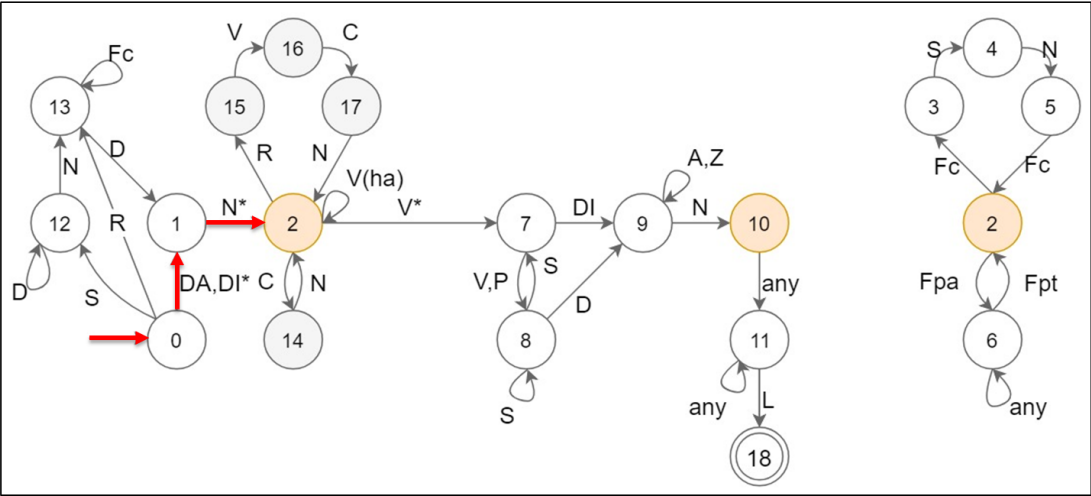


Figura 13. Flujo del autómata, parte 1

En la Figura 14 se muestra la siguiente etapa, en la que el autómata para identificar el término y género próximo recibe "N" como entrada y avanza al estado 1, el cual es un estado de control en el que el sistema transfiere el flujo al autómata principal si la siguiente entrada no conduce a otro estado. La siguiente entrada es "V". El sistema permite la transición del estado 1 al estado 2 únicamente cuando se recibe "V" como entrada y la palabra no deriva del verbo "ser". En este ejemplo, la palabra sí deriva del verbo "ser", por lo que el sistema devuelve el flujo al autómata principal.

Un *exantema* es una *erupción cutánea* generalizada que aparece como
 D N V D N A V P V C
 N A S D N A
 manifestación clínica de una enfermedad sistémica.

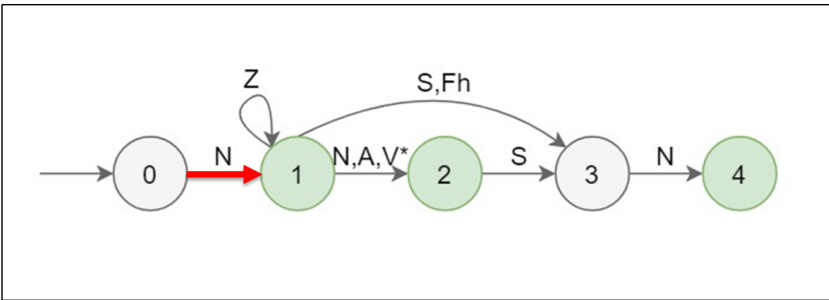


Figura 14. Flujo del autómata, parte 2

En la Figura 15, se puede observar que el sistema recibe “V” como entrada. El sistema permite la transición del estado 2 al estado 7 únicamente cuando se recibe “V” como entrada y la palabra deriva del verbo “ser”, lo cual se cumple en este ejemplo y el autómata avanza al estado 7. Después recibe “D” como entrada y avanza al estado 9. Después recibe “N” como entrada y avanza al estado 10, el cual es un estado de control en el que el sistema transfiere el flujo al autómata para identificar el término y género próximo.

Un exantema es una erupción cutánea generalizada que aparece como
 D N V D N A V P V C
manifestación clínica de una enfermedad sistémica.
 N A S D N A

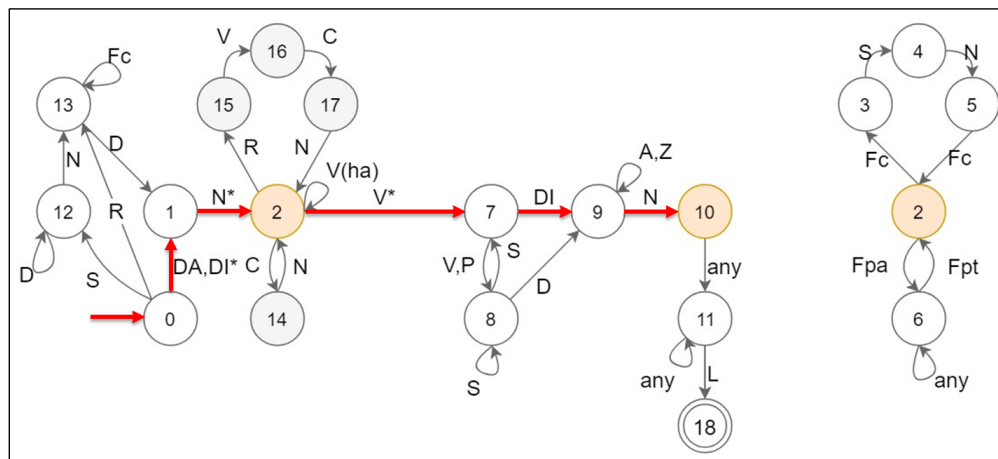


Figura 15. Flujo del autómata, parte 3

En la Figura 16, se puede observar que el sistema recibe “N” como entrada y avanza al estado 1. Después se recibe “A” como entrada y se avanza al estado 2. La siguiente entrada es “V”, la cual no conduce a ningún estado, por lo que el sistema devuelve el flujo al autómata principal.

Un *exantema* es una *erupción cutánea* generalizada que aparece como
 D N V D N A V P V C
 N A S D N A

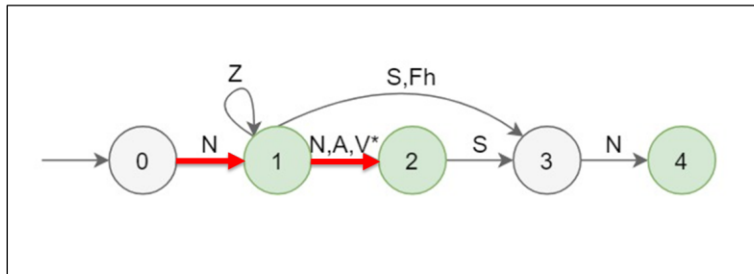


Figura 16. Flujo del autómata, parte 4

En la Figura 17 se puede observar la etapa final. A partir del estado 10, cualquier entrada conduce al estado 11. En el estado 11, cualquier entrada que no sea "L" conduce de vuelta al estado 11. Cuando se recibe "L" como entrada se avanza al estado 18, el estado final, y el sistema acepta el fragmento de texto como un contexto definitorio válido.

Un *exantema* es una *erupción cutánea* generalizada que aparece como
 D N V D N A V P V C
 N A S D N A

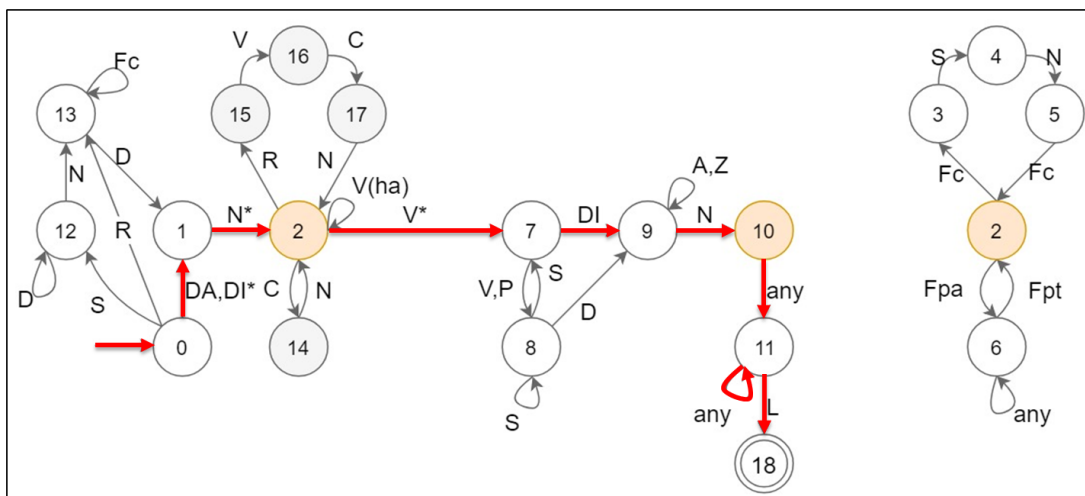


Figura 17. Flujo del autómata, parte 5

4.4. Plataforma web

Durante las investigaciones realizadas, se desarrolló una plataforma web en Ruby on Rails que permite analizar archivos de texto no estructurado e identificar los contextos definitorios y sus elementos principales. Esta plataforma se desarrolló a manera de prueba de concepto para poder visualizar de manera amigable el resultado del análisis que realiza el sistema desarrollado. En la pantalla principal hay un botón que permite seleccionar el archivo a analizar (Figura 18).

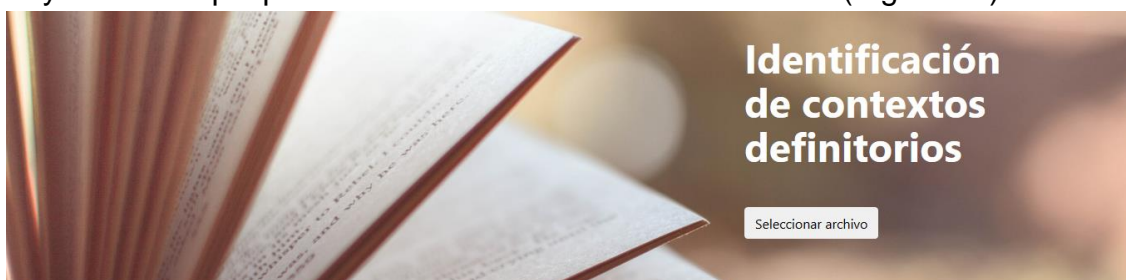


Figura 18. Pantalla principal

Éste archivo debe de ser un archivo de texto en formato “UTF-8 sin BOM”. Tras unos momentos, la plataforma presenta al usuario el texto analizado con los elementos señalados de la siguiente manera: en *negritas* los **contextos definitorios**, en color azul el **término** y en color rojo el **género próximo** (Figura 19). Adicionalmente se presenta al usuario la opción de visualizar el texto con los saltos de línea del archivo original.

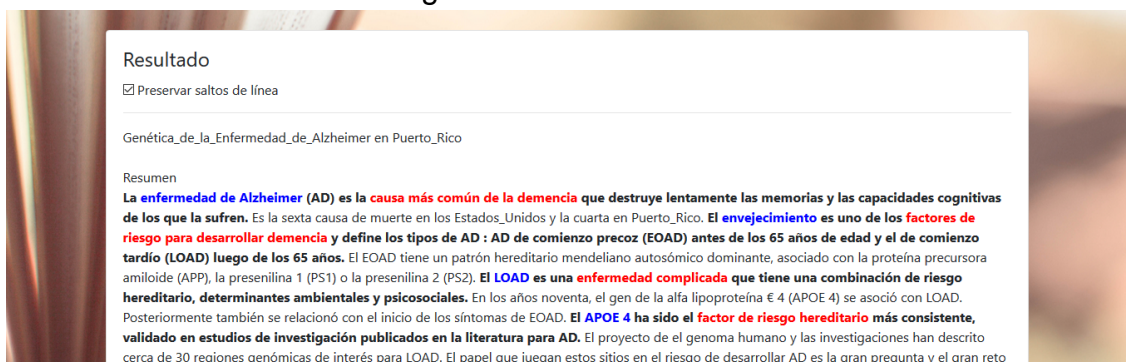


Figura 19. Pantalla de resultado

Al final del resultado se presentan 3 botones: el botón para subir al inicio de la página, el botón para analizar otro archivo, y el botón para descargar un archivo de texto con los elementos identificados (Figura 20).

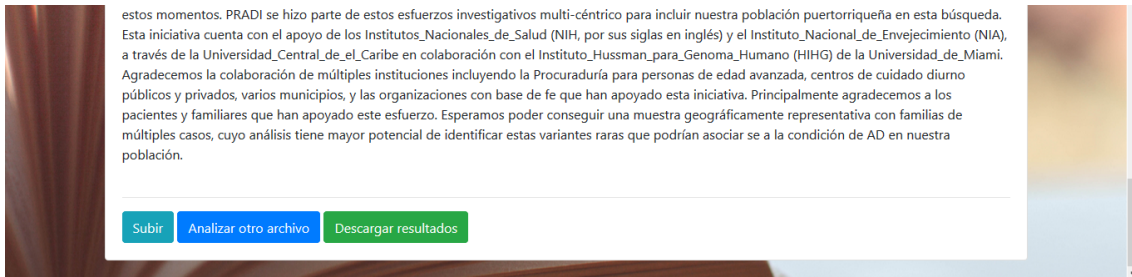


Figura 20. Opciones al pie del resultado

En la tabla 6 se muestra un ejemplo del archivo de texto generado por la aplicación al hacer clic en “Descargar resultados”, en el cual se presentan los elementos encontrados y su relación Hipónimo > Hiperónimo, descrita por el símbolo “mayor que” (>).

Tabla 11. Archivo generado por la aplicación

<p>alergia>reacción inmunitaria reacción adversa a alimentos>reacción anómala anafilaxia>reacción de hipersensibilidad dermatitis alérgica de contacto>enfermedades ocupacionales origen>respuesta alterada Aerobiología>herramienta</p>

A partir de estas relaciones es posible construir taxonomías, las cuales representan un paso importante para la creación de una ontología y para el aprendizaje ontológico.

Capítulo 5. Pruebas

Las pruebas realizadas se dividen en dos capítulos principales. El primero es la identificación de contextos definitorios, para el cual se diseñó e implementó en lenguaje Ruby un autómata capaz de identificar los diferentes patrones de contextos definitorios presentes en textos no estructurados. El segundo es la identificación del término y género próximo, para el cual se diseñó e implementó en lenguaje Ruby un autómata capaz de identificar los diferentes patrones de término y género próximo presentes en los contextos definitorios antes mencionados.

5.1. Etiquetado manual

A continuación, se describe el proceso que se llevó a cabo para el etiquetado manual de los elementos relevantes en los artículos de medicina. Vale la pena recordar que uno de los objetivos principales de este trabajo es demostrar la relación de hiperonimia que existe entre el término y el género próximo de la definición aristotélica que hace uso de verbo definitorio “ser”.

Primero se tomó como referencia la estructura del contexto definitorio propuesta en [20] la cual explica que éste se encuentra constituido por un término, una predicación verbal (PV) o marcador tipográfico (MP), una Definición y, opcionalmente, un Patrón pragmático, como se muestra en la Figura 21.

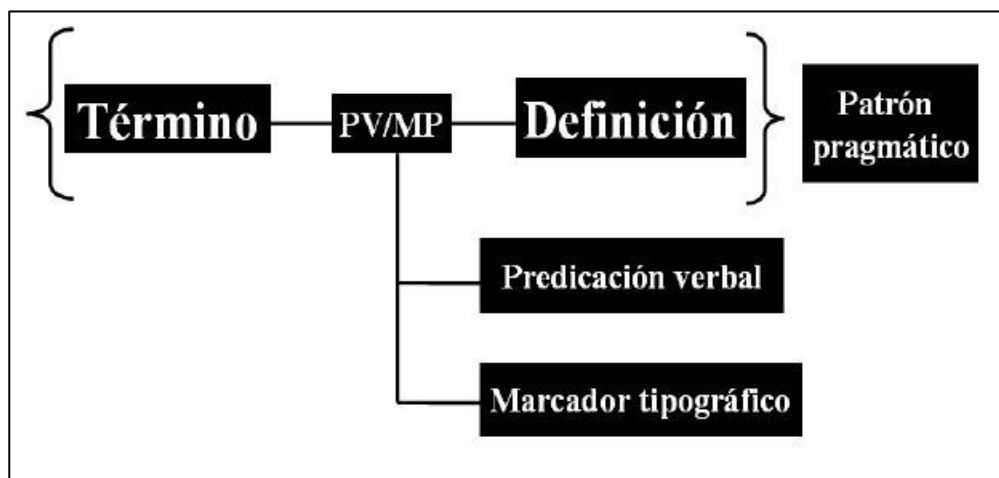


Figura 21. Elementos principales de un CD [20]

En la siguiente oración se muestra un ejemplo de la identificación manual de los elementos de un contexto definitorio, donde se resaltó en color azul el término, en rojo la predicación verbal y subrayado la definición.

*La **alergia** es una **reacción inmunitaria** del organismo frente a una sustancia generalmente inocua para el anfitrión, que se manifiesta por unos signos y síntomas característicos cuando este se expone a ella.*

En [20] también se detallan los diferentes tipos de definiciones, y se explica que la definición aristotélica consta de un género próximo y una diferencia específica. El género próximo es un término similar al que se define, mientras que la diferencia específica es un fragmento de texto que describe la diferencia entre el término y el género próximo. Haciendo uso del texto del ejemplo anterior, se muestra un nuevo ejemplo donde se resalta en color azul el término, en rojo el género próximo y subrayado la diferencia específica.

*La **alergia** es una **reacción inmunitaria** del organismo frente a una sustancia generalmente inocua para el anfitrión, que se manifiesta por unos signos y síntomas característicos cuando este se expone a ella.*

Con base a esta información, se realizó el etiquetado manual de los 20 artículos de medicina que conforman el corpus de esta investigación y se identificó un total de 50 contextos definitorios, los cuales después fueron comparados contra el etiquetado automático, como se explica en el subcapítulo de **Resultados de las pruebas**.

5.2. Identificación de contextos definitorios

Para la identificación de los contextos definitorios, se creó un script que procesa el texto mediante dos operaciones principales: El etiquetado POS del texto, realizado mediante la herramienta Freeling; y el análisis del texto a través de un autómata finito determinista.

El etiquetado POS asigna una etiqueta a cada palabra y símbolo en el texto que representa su función dentro de una oración. Adicionalmente, Freeling separa el texto por oraciones. A continuación, el autómata analiza cada etiqueta como entrada, y para cada una produce un cambio de estado y una salida.

Para probar la eficacia del programa se recopilaron 20 artículos de medicina, en los cuales se realizó el etiquetado manual de los contextos definitorios para posteriormente hacer la comparación con el etiquetado que devuelve el programa. Los documentos son artículos de una enciclopedia que contienen información sobre términos médicos, recuperados de la página https://es.wikipedia.org/wiki/Categoría:Términos_médicos [30].

A continuación, se presenta una lista de los artículos utilizados para las pruebas.

- afasia.txt
- afta.txt
- alergia.txt
- astigmatismo.txt
- botulismo.txt
- colitis.txt
- daltonismo.txt
- dislexia.txt
- ecografía.txt
- estomago.txt
- estrabismo.txt
- exantema.txt
- exodoncia.txt
- fiebre.txt
- ganglion.txt
- gingivitis.txt
- gota.txt
- hepatitis.txt
- hipertiroidismo.txt
- liposucción.txt

En la Figura 22 se presenta un fragmento del etiquetado manual de uno de los artículos utilizado para las pruebas. Los contextos definitorios fueron resaltados con la tipografía de *negritas* para diferenciarlos del resto del texto.

Un afta (de el griego antiguo ἄφθαι, aphtai, " quemaduras ") es una **úlcer**
superficial, pequeña, redondeada, blanquecina y con borde rojo bien delimitado;
de origen desconocido, que aparece durante el curso de ciertas enfermedades.
Suele ser recurrente.1 Se forma en la mucosa de la boca o de otras partes de el tubo
digestivo, o en la mucosa genital; como la presentación más habitual es la
orofaríngea, se usa con frecuencia en un sentido más restringido, referido tan solo a el
afta bucal.2 El afta bucal u oral o estomatitis aftosa o úlcera bucal es una lesión o
erosión mucosa, como una pequeña herida o llaga, que se localiza generalmente en la
mucosa oral de bordes planos y regulares y rodeada de una zona de eritema.3 Índice [
ocultar] 1 Epidemiología 2 Etiología 3 Cuadro clínico 4 Clasificación 5 Diagnóstico
diferencial 6 Tratamiento 7 Véase también 8 Referencias 9 Bibliografía 10 Enlaces
externos Epidemiología [editar] **Las aftas son una de las lesiones más frecuentes
de la cavidad bucal con una prevalencia entre el 5 y 80_% de la población.4 Se
presenta con gran frecuencia entre niños y adolescentes, especialmente entre los
10 y 19 años de edad.** Etiología [editar] Las causas que provocan el desarrollo de
las aftas bucales no están completamente claras, si bien se sabe que tienen un origen
multifactorial.

Figura 22. Etiquetado manual para la identificación de contextos definitorios

En comparación, la Figura 23 muestra cómo la plataforma web presenta los resultados de manera muy similar.

Resultado

Preservar saltos de línea

Un afta (de el griego antiguo ἄφθαι, aphtai, " quemaduras ") es una **úlcer**a superficial, pequeña, redondeada, blanquecina y con borde rojo bien delimitado; de origen desconocido, que aparece durante el curso de ciertas enfermedades. Suele ser recurrente.1 Se forma en la mucosa de la boca o de otras partes de el tubo digestivo, o en la mucosa genital; como la presentación más habitual es la orofaríngea, se usa con frecuencia en un sentido más restringido, referido tan solo a el afta bucal.2 El afta bucal u oral o estomatitis aftosa o úlcera bucal es una lesión o erosión mucosa, como una pequeña herida o llaga, que se localiza generalmente en la mucosa oral de bordes planos y regulares y rodeada de una zona de eritema.3 Índice [ocultar] 1 Epidemiología 2 Etiología 3 Cuadro clínico 4 Clasificación 5 Diagnóstico diferencial 6 Tratamiento 7 Véase también 8 Referencias 9 Bibliografía 10 Enlaces externos Epidemiología [editar] **Las aftas son una de las lesiones más frecuentes de la cavidad bucal con una prevalencia entre el 5 y 80_% de la población.**4 Se presenta con gran frecuencia entre niños y adolescentes, especialmente entre los 10 y 19 años de edad. Etiología [editar] Las causas que provocan el desarrollo de las aftas bucales no están completamente claras, si bien se sabe que tienen un origen multifactorial. Diversas alteraciones en el funcionamiento de el sistema inmunitario, tanto de origen genético (congénitas o innatas) como otras que se desarrollan a lo largo de la vida (adquiridas), juegan un importante papel.5 Entre los principales factores que modifican la respuesta inmunitaria se incluyen : deficiencias de microelementos y vitaminas, alergias a alimentos, enfermedades sistémicas (tales como la enfermedad celíaca, la enfermedad de Crohn, la colitis ulcerosa y el sida), trastornos hormonales, algunas infecciones virales y bacterianas, el aumento de el estrés oxidativo, lesiones mecánicas, la ansiedad y el

Figura 23. Resultado para la identificación de contextos definitorios

5.3. Identificación del término y género próximo

Para la identificación de estos elementos, el programa en Ruby desarrollado para la identificación de contextos definitorios cuenta con un módulo extra. Este módulo fue desarrollado a partir de un segundo autómata finito determinista, encargado exclusivamente de identificar los patrones que corresponden a un término o género próximo válido.

Para probar la eficacia del módulo, se trabajó con los mismos artículos utilizados para las pruebas de identificación de contextos definitorios y se realizó el etiquetado manual del término y género próximo presentes en cada uno de dichos contextos.

En la Figura 24 se presenta un fragmento del etiquetado manual de uno de los artículos utilizado para las pruebas. El término de cada contexto definitorio se encuentra resaltado en color azul, mientras el género próximo se encuentra resaltado en color rojo.

La liposucción es una técnica quirúrgica que se utiliza en cirugía estética y que permite un remodelado de la silueta a través de la extracción de grasa o tejido adiposo de diversos sitios de el cuerpo usando una cánula o jeringa conectada a una máquina succionadora (liposucción mecánica), o mediante ultrasonido (liposucción ultrasónica). Índice [ocultar] 1 Historia 2 Objetivos 3 Técnica 3.1 Cuidados posoperatorios 4 Complicaciones 5 Mortalidad 6 Véase también 7 Referencias 8 Enlaces externos Historia [editar] Hasta que apareció la liposucción, el exceso de grasa se trataba extirpándolo junto con la piel (dermolipectomias y abdominoplastias) dejando, en consecuencia, grandes cicatrices. Esta técnica fue inventada por el ginecólogo italiano Giorgio_Fischer en 1974; 1 sin embargo, fue el cirujano francés Gerard_Yves_Illouz el primero en utilizar la con fines estéticos en 1977.2 Illouz acopló una cánula a un aspirador e introduciéndola bajo la piel, con movimientos de vaivén, el tejido graso era desprendido y aspirado. Desde su introducción se han producido modificaciones y novedades en lo referente a cánulas, aspiradores, anestesia, y sobre todo, indicaciones más precisas. En 1985 el dermatólogo Jeffrey_Klein inventa la liposucción tumescente o método húmedo.3 A el principio esta intervención se practicaba con anestesia general.

Figura 24. Etiquetado manual para la identificación del término y género próximo

La plataforma web desarrollada para probar el funcionamiento del programa presenta los resultados de manera muy similar, como se puede ver en la Figura 25.

Resultado

Preservar saltos de línea

La liposucción es una técnica quirúrgica que se utiliza en cirugía estética y que permite un remodelado de la silueta a través de la extracción de grasa o tejido adiposo de diversos sitios de el cuerpo usando una cánula o jeringa conectada a una máquina succionadora (liposucción mecánica), o mediante ultrasonido (liposucción ultrasónica). Índice [ocultar] 1 Historia 2 Objetivos 3 Técnica 3.1 Cuidados posoperatorios 4 Complicaciones 5 Mortalidad 6 Véase también 7 Referencias 8 Enlaces externos Historia [editar] Hasta que apareció la liposucción, el exceso de grasa se trataba extirpándolo junto con la piel (dermolipectomias y abdominoplastias) dejando, en consecuencia, grandes cicatrices. Esta técnica fue inventada por el ginecólogo italiano Giorgio_Fischer en 1974; 1 sin embargo, fue el cirujano francés Gerard_Yves_Illouz el primero en utilizar la con fines estéticos en 1977.2 Illouz acopló una cánula a un aspirador e introduciéndola bajo la piel, con movimientos de vaivén, el tejido graso era desprendido y aspirado. Desde su introducción se han producido modificaciones y novedades en lo referente a cánulas, aspiradores, anestesia, y sobre todo, indicaciones más precisas. En 1985 el dermatólogo Jeffrey_Klein inventa la liposucción tumescente o método húmedo.3 A el principio esta intervención se practicaba con anestesia general. La liposucción era realizada sólo en las porciones más profundas de la grasa y se empleaban cánulas de gran diámetro. Si el cirujano se acercaba excesivamente a la piel existía, debido a el tamaño de las cánulas, un alto riesgo de irregularidades. La intervención era muy traumática. Actualmente la mayoría de las intervenciones

Figura 25. Resultado para la identificación del término y género próximo

5.4. Resultados de las pruebas

5.4.1. Identificación de contextos definitorios

En los 20 archivos analizados, se encontraron 50 contextos definitorios en el etiquetado manual, de los cuales el sistema fue capaz de identificar 37 correctamente.

Para calcular la eficacia del programa, se comparó el etiquetado manual con el etiquetado que presenta el programa y se obtuvo el valor de precisión y cobertura para cada uno de los textos analizados. La fórmula para calcular estos valores es la siguiente:

$$\text{Precisión} = VP / (VP + FP)$$

$$\text{Cobertura} = VP / (VP + FN)$$

Los resultados de precisión y cobertura para la identificación de contextos definitorios se presentan en la siguiente tabla.

Tabla 12. Precisión y cobertura para la identificación de contextos definitorios

Archivo	Verdadero		Falso		Precisión	Cobertura
	Positivo	Negativo	Positivo	Negativo		
afasia	0	0	1	2	0.00%	0.00%
afta	2	0	0	1	100.00%	66.67%
alergia	5	0	0	4	100.00%	55.56%
astigmatismo	1	0	0	0	100.00%	100.00%
botulismo	3	0	0	0	100.00%	100.00%
colitis	2	0	1	2	66.67%	50.00%
daltonismo	2	0	1	1	66.67%	66.67%
diselxia	2	0	0	0	100.00%	100.00%
ecografia	2	0	0	0	100.00%	100.00%
estomago	2	0	0	0	100.00%	100.00%
estrabismo	1	0	0	1	100.00%	50.00%
exantema	2	0	0	1	100.00%	66.67%
exodoncia	2	0	0	0	100.00%	100.00%
fiebre	1	0	0	0	100.00%	100.00%
ganglion	2	0	0	0	100.00%	100.00%
gingivitis	1	0	0	0	100.00%	100.00%
gota	2	0	1	0	66.67%	100.00%
hepatitis	2	0	0	1	100.00%	66.67%
hipertiroidismo	2	0	0	0	100.00%	100.00%
liposucción	1	0	0	0	100.00%	100.00%
Total					90.24%	74.00%

El sistema obtuvo un resultado de 90.24% en precisión y 74% en cobertura. No se logró conseguir un mayor nivel de cobertura debido a que existen múltiples excepciones en los patrones verbales que deben ser analizadas con detalle en futuras investigaciones. Estas excepciones se describen con mayor detalle en el capítulo de **Limitaciones detectadas**.

5.4.2. Identificación del término y género próximo

Partiendo de los mismos archivos, en los 50 contextos definitorios se etiquetaron manualmente 50 términos y 50 género próximo, de los cuales el sistema fue capaz de identificar 73 de manera correcta y 1 de manera incorrecta, donde el término multipalabra identificado incluía una palabra más de lo esperado.

Los resultados de la comparación entre el etiquetado manual y el etiquetado que presenta el programa se presentan en la siguiente tabla.

Tabla 13. Eficacia en la identificación del término y género próximo

Archivo	Correcta	Incorrecta
afasia	0	0
afta	4	0
alergia	10	0
astigmatismo	2	0
botulismo	6	0
colitis	4	0
daltonismo	4	0
diselxia	4	0
ecografia	4	0
estomago	4	0
estrabismo	2	0
exantema	4	0
exodoncia	4	0
fiebre	2	0
ganglion	4	0
gingivitis	2	0
gota	3	1
hepatitis	4	0
hipertiroidismo	4	0
liposuccion	2	0

5.5. Limitaciones detectadas

A continuación, se describen los casos en los que la estructura del texto (o algún otro factor) presenta conflictos con los patrones definidos, afectando la correcta identificación de los elementos buscados.

5.5.1. Identificación de contextos definitorios

Determinante entre el verbo y el género próximo

Para la mayoría de los casos, una definición analítica en la que el término y género próximo presentan una relación de hiperonimia, el elemento entre el verbo y el género próximo debe ser un determinante indefinido. Sin embargo, hay casos en los que el determinante es un artículo o simplemente no se encuentra. Por ejemplo, en el caso de la oración:

*La **afasia** es **el** **trastorno** del lenguaje que se produce como consecuencia de una patología cerebral.*

Se produce un falso negativo ya que el determinante es un artículo. Se puede observar un error similar en el siguiente ejemplo:

*Las **petequias** son **lesiones puntiformes** de color rojo púrpura, por extravasación de sangre que no desaparecen con la **dígito-presión**.*

En el cual se produce un falso negativo ya que no se presenta un determinante entre el verbo y el género próximo.

Patrón pragmático

Las definiciones a veces contienen información sobre el contexto en el cual se describe un término. Por ejemplo, en el caso de la oración:

*La **difasia** **por otro lado** es un **trastorno** específico en la adquisición del lenguaje.*

Se produce un falso negativo pues el patrón pragmático “por otro lado” se encuentra entre el término y el verbo definitorio, interrumpiendo el flujo del autómata.

Error en el etiquetado

Hay ciertas palabras que el etiquetado POS no logra identificar correctamente, lo cual afecta el flujo del autómata. Por ejemplo, en el caso de la siguiente oración:

*La **tritanopia** es una **condición** muy poco frecuente en la que están ausentes los fotorreceptores de la retina para el color azul.*

Se produce un falso negativo, ya que la palabra “tritanopia” es etiquetada como un adjetivo.

No hay definición

Hay casos en los que el patrón se cumple, pero la oración no corresponde con una definición o los elementos identificados no presentan una relación de hiperonimia. Por ejemplo, en el caso de la oración:

...se distingue de la alergia en que estas últimas el origen es una respuesta alterada de el sistema inmune...

Se produce un falso positivo ya que, aunque se detecta la estructura de un CD, el texto no corresponde con una definición.

5.5.2. Identificación de término y género próximo

Adjetivo descriptor

En algunas ocasiones, el adjetivo no describe al sustantivo, sino que da inicio a la diferencia específica. Por ejemplo, en el caso de la siguiente oración:

La gota es una enfermedad producida por una acumulación de cristales de urato monosódico...

La palabra “producida” forma el término multpalabra “enfermedad producida”, el cual no puede considerarse como una entidad en una ontología.

Las limitaciones mencionadas fueron detalladas en este capítulo para llevar a cabo una investigación adicional en trabajos futuros, que permitan mejorar los resultados de precisión y cobertura del sistema desarrollado.

Capítulo 6. Conclusiones y trabajos futuros

El programa desarrollado es capaz de analizar textos no estructurados para identificar los contextos definitorios que contienen una definición analítica. Así mismo, en estas definiciones se identifica el término y el género próximo. El programa ofrece resultados con un índice de 90.24% en precisión y 74% en cobertura. Esta herramienta puede agilizar la tarea de estructurar la información existente en los medios digitales que se utilizan diariamente, mediante la creación de ontologías a partir de algunos de los términos más relevantes de esta información.

Un área de estudio que puede hacer uso de esta herramienta es la del Aprendizaje ontológico, la cual se enfoca en la detección de los elementos principales de un tema de discurso y la relación que existe entre estos para la creación automática de ontologías. La aportación de nuestra herramienta es la identificación de algunos de los elementos principales de un texto (los conceptos principales de una definición) y la relación de hiperonimia existente entre ellos.

Durante el desarrollo de la investigación y las pruebas realizadas, se pudo observar que existen diferentes relaciones semánticas entre los elementos principales de un contexto definitorio, como la relación de hiperonimia, sinonimia e individuación. Es posible etiquetar los tipos de relaciones mencionados haciendo estudios adicionales sobre los diferentes patrones que presentan los contextos definitorios. De ser posible, esto facilitará aún más la tarea de creación y poblado automático de ontologías.

Como trabajos futuros se tiene lo siguiente:

- 1) *Realizar pruebas con corpus de un dominio distinto al de medicina, para evaluar el desempeño del sistema desarrollado en otros dominios.*
- 2) *Refinar los patrones definitorios para la identificación de contextos definitorios que contienen una definición analítica, para así alcanzar un mayor índice de precisión y cobertura en la identificación de estos elementos.*
- 3) *Crear y poblar ontologías a partir de las taxonomías encontradas en los contextos definitorios que contienen una definición analítica, para poder visualizar y evaluar el beneficio de esta técnica de aprendizaje ontológico en la estructuración de la información.*
- 4) *Investigar e implementar patrones definitorios para la identificación de contextos definitorios que contienen una definición no analítica, los cuales podrían describir otro tipo de relaciones semánticas.*
- 5) *Diseñar e implementar un método que haga uso de otras técnicas de análisis de textos, como las expresiones regulares.*

Referencias

- [1] J. A. Reyes Ortiz, "Creación Automática de Ontologías a partir de Textos."
- [2] J. D. G. Fierros, "Poblado automático de ontologías espaciales." 2012.
- [3] G. Sierra, R. Alarcón, C. Aguilar, and C. Bach, "Definitional verbal patterns for semantic relation extraction," *Terminology*, vol. 14, no. 2008, pp. 74–98, 2008.
- [4] R. Alarcón, "Extracción automática de contextos definitorios en corpus especializados," Universidad Pompeu Fabra, Barcelona, 2009.
- [5] R. Alarcón and G. Sierra, "The Role of Verbal Predications for Definitional Contexts Extraction," 2003.
- [6] R. Alarcón and G. Sierra, "Reglas lexico-metalinguísticas para la extracción de automática de contextos definitorios," *Av. en la Cienc. la Comput.*, pp. 242–247, 2006.
- [7] R. Alarcón, G. Sierra, and C. Bach, "Developing a Definitional Knowledge Extraction System," *Evaluation*, pp. 2–6.
- [8] V. Kumar Bhatia, P. Sánchez Hernández, and P. Pérez-Paredes, *Researching Specialized Languages*. 2011.
- [9] N. Aussenac-Gilles, B. Biébow, and S. Szulman, "D'une méthode à un guide pratique de modélisation de connaissances à partir de textes," *TIA-2003 - Actes des cinquièmes rencontres Terminol. Intell. Artif.*, pp. 41–53, 2003.
- [10] Y. Marrero García, P. Moreda Pozo, and R. Muñoz-Guillena, "Pattern Construction for Extracting Domain Terminology," *Proc. Int. Conf. Recent Adv. Nat. Lang. Process.*, pp. 420–426, 2015.
- [11] W. IJntema, J. Sangers, F. Hogenboom, and F. Frasincar, "A lexico-semantic pattern language for learning ontology instances from text," *J. Web Semant.*, vol. 15, pp. 37–50, 2012.
- [12] G. Sierra, "Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos," 2009.
- [13] C. Aguilar, "Análisis lingüístico de definiciones en contextos definitorios," no. November, 2009.
- [14] C. Aguilar, R. Alarcón, C. Rodríguez, and G. Sierra, "Reconocimiento y clasificación de patrones verbales definitorios en corpus especializados.," *La Terminol. en el siglo XXI Contrib. a la Cult. la paz, la Divers. y la sostenibilidad*.
- [15] C. Aguilar and G. Sierra, "Hacia una tipología de definiciones basada en el modelo analítico," in *XV Congreso Internacional ALFAL*, 2008.
- [16] M. M. Suarez de la Torre, L. F. Castillo Ossa, C. Ríos Cardona, G. M. Muñoz, and J. Aranzazu Álvarez, "Análisis , diseño e implementación de un agente deliberativo para extraer contextos definitorios en textos especializados," *Rev. Interam. Bibl.*, vol. 32, no. 2, pp. 59–84, 2009.
- [17] D. Freitag, "Machine learning for information extraction in informal domains," *Mach. Learn.*, vol. 39, no. 2–3, pp. 169–202, 2000.
- [18] R. A. Española, *Diccionario de la lengua española*. Madrid, 2001.
- [19] A. Copestake, *Inheritance, defaults and the lexicon*. New York, 1993.
- [20] M. A. Dorantes Cruz, "Sintaxis de la definición analítica para sustantivos en un diccionario especializado," Facultad de estudios superiores

- Acatlán, 2016.
- [21] P. Lluís and E. Stanilovsky, "FreeLing 3.0: Towards Wider Multilinguality," in *Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA*, 2012.
 - [22] L. Padró, "Analizadores multilingües en freeling," *Linguamática*, vol. 3, pp. 13–20, 2012.
 - [23] L. Padró, M. Collado, S. Reese, M. Lloberes, and I. Castellón, "FreeLing 2.1: Five years of open-source language processing tools," *English*, pp. 931–936, 2002.
 - [24] J. Atserias, B. Casas, E. Comelles, M. González, L. Padró, and M. Padró, "FreeLing 1.3: Syntactic and semantic services in an open-source NLP library," *Proc. 5th Int. Conf. Lang. Resour. Eval. Lr.*, no. February 2017, pp. 48–55, 2006.
 - [25] X. Carreras, I. Chao, L. Padró, and M. Padró, "Freeling: An Open-Source Suite of Language Analyzers," *Proc. 4th Lang. Resour. Eval. Conf. (LREC 2004)*, vol. 4, pp. 239–242, 2004.
 - [26] G. Leech and A. Wilson, *Recommendations for the Morphosyntactic Annotation of Corpora*. Lancaster, 1996.
 - [27] Merriam-Webster.com, "ontology.," 2017. [Online]. Available: <https://www.merriam-webster.com>. [Accessed: 24-Aug-2017].
 - [28] P. Buitelaar and P. Cimiano, *Ontology learning and population: bridging the gap between text and knowledge*, vol. 167. 2008.
 - [29] C. Aguilar, "Análisis lingüístico de definiciones en contextos definatorios," no. November, 2009.
 - [30] C. de Wikipedia, "Términos médicos." [Online]. Available: https://es.wikipedia.org/wiki/Categoría:Términos_médicos. [Accessed: 10-Jan-2018].

Anexos

A continuación, se presenta un fragmento de algunos de los artículos utilizados para el etiquetado de contextos definitorios. Los archivos originales son más extensivos, abarcando de 3 a 15 hojas.

afasia.txt

La afasia es el trastorno del lenguaje que se produce como consecuencia de una patología cerebral.

Se trata de la pérdida de capacidad de producir o comprender el lenguaje, debido a lesiones en áreas cerebrales especializadas en estas funciones. Es entonces una pérdida adquirida en el lenguaje oral. 2 El término afasia, que fue creado en 1864 por el médico francés Armand Trousseau (1801-1867), procede del vocablo griego ἀφασία, 'imposibilidad de hablar'. La disfasia por otro lado es un trastorno específico en la adquisición del lenguaje.

El hemisferio cerebral izquierdo como base del lenguaje[editar]

Aunque de apariencia similar, los hemisferios cerebrales se especializan en funciones diferentes. Una de las más conocidas es la especialización del hemisferio izquierdo en la mayoría de las personas como base del lenguaje. La comprensión de los aspectos no verbales del lenguaje y de la prosodia de éste se relacionan más directamente con el hemisferio derecho. Esto es así para el 98,5% de las personas diestras y el 70% de las personas zurdas, estando parcial o totalmente lateralizado en el lado derecho en el resto.

El hemisferio izquierdo también se encarga de controlar la motricidad de los miembros del hemicuerpo (mitad del cuerpo) derecho. Además, las zonas motoras se encuentran físicamente cercanas a las del lenguaje, por lo que es común que algunos subtipos de afasia se acompañen de hemiparesia (debilidad motora en un hemicuerpo).

Etiología[editar]

La afasia puede ser causada por un accidente cerebrovascular, un traumatismo craneoencefálico, una infección cerebral, una neoplasia o un proceso degenerativo.

- Accidente cerebrovascular: es la causa más frecuente de afasia, sobre todo el producido por isquemia trombótica o embolígena.
- Traumatismo craneoencefálico: provocado generalmente por un accidente.
- Infecciones localizadas o difusas del cerebro, como absceso cerebral o encefalitis.
- Tumores del Sistema Nervioso Central.
- Enfermedades degenerativas, como la enfermedad de Alzheimer o la Enfermedad de Parkinson.

Clasificación[editar]

Existen dos formas básicas de afasia: la afasia de Broca y la afasia de Wernicke6

[...]

afta.txt

Un afta (del griego antiguo ἄφθαι, aphtai, "quemaduras") es una úlcera superficial, pequeña, redondeada, blanquecina y con borde rojo bien delimitado; de origen desconocido, que aparece durante el curso de ciertas enfermedades. Suele ser recurrente.¹ Se forma en la mucosa de la boca o de otras partes del tubo digestivo, o en la mucosa genital; como la presentación más habitual es la orofaríngea, se usa con frecuencia en un sentido más restringido, referido tan solo al afta bucal.²

El afta bucal u oral o estomatitis aftosa o úlcera bucal es una lesión o erosión mucosa, como una pequeña herida o llaga, que se localiza generalmente en la mucosa oral de bordes planos y regulares y rodeada de una zona de eritema.³

Índice [ocultar]

- 1 Epidemiología
- 2 Etiología
- 3 Cuadro clínico
- 4 Clasificación
- 5 Diagnóstico diferencial
- 6 Tratamiento
- 7 Véase también
- 8 Referencias
- 9 Bibliografía
- 10 Enlaces externos

Epidemiología[editar]

Las aftas son una de las lesiones más frecuentes de la cavidad bucal con una prevalencia entre el 5 y 80% de la población.⁴ Se presenta con gran frecuencia entre niños y adolescentes, especialmente entre los 10 y 19 años de edad.

Etiología[editar]

Las causas que provocan el desarrollo de las aftas bucales no están completamente claras, si bien se sabe que tienen un origen multifactorial. Diversas alteraciones en el funcionamiento del sistema inmunitario, tanto de origen genético (congénitas o innatas) como otras que se desarrollan a lo largo de la vida (adquiridas), juegan un importante papel.⁵

Entre los principales factores que modifican la respuesta inmunitaria se incluyen: deficiencias de microelementos y vitaminas, alergias a alimentos, enfermedades sistémicas (tales como la enfermedad celíaca, la enfermedad de Crohn, la colitis ulcerosa y el sida), trastornos hormonales, algunas infecciones virales y bacterianas, el aumento del estrés oxidativo, lesiones mecánicas, la ansiedad y el tabaco.⁶⁵

[...]

alergia.txt

La alergia es una reacción inmunitaria del organismo frente a una sustancia generalmente inocua para el anfitrión, que se manifiesta por unos signos y Síntomas característicos cuando este se expone a ella (por inhalación, ingestión o contacto cutáneo). Durante mucho tiempo la alergia se ha considerado equivalente a la hipersensibilidad (un término más antiguo) y por ello se ha considerado erróneamente como una reacción inmunitaria exagerada ante una sustancia. Pero la «alergia» es la expresión clínica de los mecanismos de respuesta inmunitarios normales del organismo, frente a los posibles invasores; y el error no está en el tipo de respuesta ni en su intensidad sino en el objetivo, que no constituye ninguna amenaza. La consecuencia final de este error del sistema inmunitario es la enfermedad del anfitrión, provocada por los efectos colaterales sufridos por los tejidos, allí donde el sistema inmunitario trata de defenderse de esa sustancia inocua. Las manifestaciones clínicas de esta enfermedad son diversas, ya que dependen de la sustancia causal y del órgano afectado. En la actualidad, más de un tercio de la población mundial presenta alguna enfermedad de origen alérgico.¹

La alergia es la causa fundamental de enfermedades tan frecuentes como la conjuntivitis, la rinitis o el asma y de enfermedades tan graves como la anafilaxia. Desde hace casi 100 años, la alergología es la especialidad médica que se encarga del estudio, diagnóstico y tratamiento de este grupo de enfermedades y los profesionales médicos que la desempeñan se denominan alergólogos (España) o alergistas (Sudamérica).

Problemas terminológicos[editar]

El término «alergia» se acuñó en respuesta a una nueva concepción del sistema inmunitario y como una solución a un problema terminológico (v. sección Historia) pero, paradójicamente, ha suscitado a lo largo de su historia muchos problemas conceptuales que aún hoy persisten y han dotado al término de ambigüedad.

En octubre de 2003, el Comité de Revisión de Nomenclatura de la Organización Mundial de Alergia (WAO)² actualizó la declaración de consenso, de expertos sobre terminología en alergia, publicada en 2001 por la Academia Europea de Alergología e Inmunología Clínica (EAACI),³ con el objetivo de acabar con la ambigüedad. El informe definió primero el término amplio de «hipersensibilidad» como «los síntomas o signos objetivamente reproducibles iniciados por la exposición a un estímulo definido a una dosis tolerada por personas normales». La hipersensibilidad podía dividirse en un tipo no alérgico, cuando no podía demostrarse ningún mecanismo inmunitario, y otro alérgico, cuando sí podía demostrarse tal mecanismo. Por lo tanto, es muy importante consignar siempre, a la hora de referirnos a este tipo de reacciones, el mecanismo implicado: anticuerpos IgE, IgG, IgM, inmunocomplejos o celular). Las enfermedades clásicas consideradas alérgicas son las mediadas por anticuerpos IgE, y por ello hoy deberíamos escribir siempre al consignarlas «hipersensibilidad alérgica mediada por IgE» mejor que simplemente «alergia».

[...]