# Metodología para la Resolución de Topónimos basada en Ontologías con Razonamiento Espacial


Angeles Belém Priego Sánchez

Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla

Tesis para obtener el título de:
Maestra en Ciencias de la Computación

Asesor
Dra. María Josefa Somodevilla García

Puebla, México Noviembre, 2012

# Resumen

Internet y la World Wide Web se han convertido en un enorme repositorio de información consultado diariamente por millones de usuarios. Además, otros repositorios de información, como las bases de datos documentales o las bibliotecas digitales, también han aumentado su popularidad considerablemente. La mayoría de la información está compuesta por documentos textuales no estructurados y las referencias geográficas se dan por medio de nombres de lugares (topónimos). Por este motivo, este trabajo se centra principalmente en la investigación y desarrollo de una metodología para la tarea de Desambiguación de Topónimos para el idioma Español, debido a que constituye una de las tareas de la Recuperación de Información Geográfica y resaltando que existen pocas investigaciones para este idioma realizadas sobre el tema en comparación con otros idiomas, específicamente con el idioma Inglés. La metodología consiste en la utilización de una ontología como repositorio de sentidos, un corpus enriquecido como contexto y un método desambiguador basado en un modelo de clasificación. La ontología construida a lo largo del desarrollo de esta investigación describe el espacio geográfico de la República Mexicana considerando la representación de objetos geográficos naturales y artificiales. El uso de ontologías ayuda a resolver problemas tales como encontrar el verdadero sentido de las palabras, incluyendo en un solo repositorio todos los conceptos y relaciones del dominio de trabajo. De esta manera se evitan errores en el manejo de la información y es posible unificar el lenguaje de la comunicación en función de sus diferentes sentidos semánticos a desambiguar. Se plantea utilizar un método basado en corpus, debido a que existen evidencias que indican que este tipo de métodos tienden a ser más precisos que los basados en conocimiento. El método desambiguador está basado en modelos de clasificación y la evaluación de éste muestra resultados favorables.

# Agradecimientos

Un agradecimiento especial para mi mami Normita y mi abuelita Conchita,
a quienes agradezco de todo corazón por su amor, cariño y comprensión.
Las amo y saben que son mi fortaleza.

A mis hermanos Edwins, Diego y Abihú,
por su amor y cariño, en las épocas de adversidad estamos de pie.

A mis familiares por la compañía y apoyo que me brindan,
se que cuento con ustedes siempre.

A mi asesora la Dra. María Josefa Somodevilla García,
por sus conocimientos, paciencia y consejos científicos
durante el desarrollo de esta tesis.

Al comité formado por el Dr. Ivo Pineda, el Dr. David Pinto,
la Dra. Concepción Pérez y la Dra. Darnés Vilariño,
quienes me ayudaron con sus conocimientos y profesionalismo.

A los profesores que con su fervor y pasión académica
me contagiaron y marcaron mi vida profesional.

A aquellas personas que han sido
mi soporte y compañía durante este periodo de estudio.
Una de ellas eres tú pk.

A todos muchas gracias...

# Dedicatoria

Dedico este trabajo de tesis a mi mamá, mi abuelita,
mis hermanos y familiares,
quienes me han apoyado
para cumplir este sueño académico.

Y a todas las personas
que han estado conmigo durante todo este tiempo...

*Mon rêve commence maintenant . . .*

Computers are magnificent tools for the realization of our dreams,

but no machine can replace the human spark of spirit,

compassion, love, and understanding.

Louis Gerstner

# Índice general

# Índice de Figuras

# Índice de Tablas

# CAPÍTULO 1. INTRODUCCIÓN

En este capítulo se reseña la contextualización y motivación, así como también los objetivos planteados y el alcance e interés de esta tesis. Finalmente se describe la organización del presente trabajo.

# Capítulo 1

# Introducción

Esta tesis presenta los resultados del trabajo que se han realizado en el ámbito de un campo reciente de investigación denominado recuperación de información geográfica, en específico se aborda la tarea de desambiguación de topónimos. Como se mostrará a lo largo de este documento, las principales aportaciones son una ontología espacial que describe el espacio geográfico de la República Mexicana considerando los objetos geográficos naturales y artificiales y una arquitectura completa para la desambiguación de topónimos que utiliza dicha ontología espacial y un corpus. Estos últimos son los pilares fundamentales del proceso, para de este modo obtener un modelo de clasificación que es capaz de desambiguar aplicando diferentes clasificadores y teniendo en cuenta las características para este tipo de sistemas.

## 1.1. Contextualización y motivación

Internet y la *World Wide Web* se han convertido en un enorme repositorio de información consultado diariamente por millones de usuarios. Además, otros repositorios de información, como las bases de datos documentales o las bibliotecas digitales, también han aumentado su popularidad considerablemente. Esto ha provocado que la Recuperación de Información (*Information Retrieval, IR*) se haya convertido en una de las áreas de investigación más importantes dentro de la informática y que recientemente haya experimentado un desarrollo espectacular motivado por el crecimiento de Internet y la necesidad de realizar búsquedas en la Web. Una característica importante de la IR es que se ocupa de los problemas de recuperar información por su contenido y no por sus metadatos por lo que existen técnicas para recuperar información de diversos tipos: textos, imágenes, archivos de sonido y video, entre otros.

Aunque estos repositorios contienen información de distinta naturaleza, la información más habitual es de tipo textual. A menudo, en el texto de un documento se pueden encontrar *referencias geográficas* que permiten asignar a ese documento una zona del espacio en la cual es relevante. El tener en cuenta estas referencias geográficas proporciona un valor añadido a los sistemas de recuperación de información clásicos. Los usuarios de los

sistemas demandan cada vez más servicios que les permitan situar la información recuperada en un mapa. Además, también está aumentando el interés en consultas que permitan recuperar documentos relevantes no sólo para un tema determinado sino también para una zona determinada. Es importante que los sistemas informáticos sean capaces de extraer y procesar información geográfica contenida en textos electrónicos. La mayor parte de este tipo de información está formada por nombres de lugares, llamados también topónimos.

La necesidad de gestionar esta información ha sido uno de los factores clave en la consolidación de la informática; el desarrollo de arquitecturas de sistemas, estructuras de indexación y otros componentes que permitan satisfacer estas necesidades es el objetivo principal de una nueva área de investigación denominada Recuperación de Información Geográfica (*Geographical Information Retrieval, GIR*).

La ambigüedad de los topónimos constituye un problema importante en la tarea GIR, dado que en esta tarea las peticiones de los usuarios están vinculadas geográficamente. Ha habido un gran esfuerzo por parte de la comunidad de investigadores para encontrar métodos de IR específicos para GIR que sean capaces de obtener resultados mejores que las técnicas tradicionales de IR. El campo de investigación de la GIR es aún joven, SPIRIT (*Spatially- Aware Information Retrieval on the Internet*) se puede considerar el primer proyecto importante en el área y las publicaciones derivadas de él son un buen punto de partida para introducirse en el tema ([1], [2], [3], [4]). Los trabajos posteriores proponen mejoras de los distintos componentes del sistema, fundamentalmente de su estructura de indexación.

La ambigüedad de los topónimos es probablemente un factor muy importante en la incapacidad de los sistemas GIR actuales por conseguir una ventaja a través del procesamiento de las informaciones geográficas [5]. En los últimos años, se ha experimentado el resurgimiento de un modelo empirista del tratamiento de la lengua. La materia prima para este modelo la constituyen grandes volúmenes de información textual no restringida mejor conocida como corpus. Con este resurgimiento aparece también la necesidad de desarrollar técnicas para facilitar el tratamiento de estos grandes volúmenes de datos.

La desambiguación de topónimos (DT) es una de las tareas de la IR, más específicamente de GIR y búsquedas de respuestas (*Question Answering, QA*), incluyendo la generación de mapas. Tiene como objetivo relacionar nombres de lugares con su representación geográfica. La Geo-información es abundante en la Web y bibliotecas digitales, por ejemplo, en colecciones de fotografías geo-referenciadas (Flickr), noticias, bases de datos de información demográfica (en México a cargo del Instituto Nacional de Estadística, Geografía e Informática, INEGI), entre otras. Se ha reportado que aproximadamente el 80% de las páginas web contienen referencias a lugares [6]; mucha de la información necesaria está relacionada a un contexto geográfico dado, por ejemplo, encontrar los restaurantes más cercanos, encontrar noticias acerca de un cierto país, así como encontrar fotografías tomadas en alguna comunidad en específico, entre otras.

Se ha documentado además que aproximadamente, el 20% de las consultas en la Web tienen un componente geográfico, lo cual muestra la importancia de trabajar con técnicas para el tratamiento de cierta terminología como es el caso de los topónimos. La GIR pertenece a una rama de la recuperación de información, e incluye todas las tareas de investigación que tradicionalmente forman el núcleo de la IR, pero además con un ´énfasis en la información Geográfica y Espacial.

En la tarea de GIR, la mayoría de las peticiones de los usuarios son del tipo *X* en *P* donde *P* representa un nombre de lugar y *X*, la parte temática de la consulta. GIR aborda dificultades de IR [2], tales como: ambigüedad geográfica (topónimos), por ejemplo: existe una catedral en St. Paul en Londres y otra en Sao Paulo, regiones geográficas mal definidas: "cerca del este", regiones geográficas complejas: "cerca de ciudades rusas" o "a lo largo de la costa mediterránea", aspectos multilingües: "GreaterLisbon" en inglés es lo mismo que "Grande Lisboa" en portugués o que "GroBraumLissabon" en alemán y la granularidad en las referencias a países: "al norte de Italia".

Los topónimos (nombres de lugares) pueden ser ambiguos y pueden tener alguno de los dos tipos de ambigüedad GEO/GEO o GEO/NO-GEO. La ambigüedad de topónimos tipo GEO/GEO tiene la característica principal de que un topónimo representa varios lugares; por ejemplo: "Trípoli" que es el nombre de 16 lugares. El caso de la ambigüedad de topónimos tipo GEO/NO-GEO tiene la peculiaridad primordial de que un topónimo se refiere a entidades (lugares) y no geográficas (personas, organizaciones, entre otras); por ejemplo: "Benito Juárez" representa una persona y lugares, "Java" es una isla Indonesia y un lenguaje de programación. Para ambos casos, los dominios de aplicación son la extracción y recuperación de información.

En este trabajo, se aborda la desambiguación de topónimos, específicamente la ambigüedad tipo GEO/NO-GEO, debido a que considera características especiales asociadas estrechamente a la naturaleza espacial de la información con la que se trabaja. Para este propósito se plantea usar un método basado en corpus para la tarea de desambiguación de topónimos, debido a que existen evidencias que indican que este tipo de métodos tienden a ser más precisos que los basados en conocimiento [7,8].

Por lo tanto, este trabajo se centra principalmente en la investigación y desarrollo de una metodología para la desambiguación de topónimos utilizando una ontología como repositorio de sentidos, como contexto un corpus y enriqueciéndolo, para que el método desambiguador se base en un modelo de clasificación.

## 1.2. Objetivos de la investigación

Para el desarrollo de esta tesis se plantea el siguiente objetivo general:

Desarrollar una metodología para la resolución de topónimos ambiguos para el idioma Español; utilizando una ontología como repositorio de sentidos, un corpus enriquecido como contexto y un método desambiguador basado en un modelo de clasificación.

Los objetivos particulares del presente trabajo de tesis se especifican a continuación:

1. Modelar una ontología espacial de topónimos en Español con clases de objetos geográficos naturales y artificiales. Con el fin de organizar la distribución del espacio geográfico de la República Mexicana.

2. Preprocesar recursos léxicos para enriquecer los contextos utilizados en el proceso de desambiguación.

3. Desarrollar un método desambiguador que considerando las nuevas clases y slots de la ontología extendida, realice la desambiguación en base a un modelo de clasificación.

4. Obtener un modelo de clasificación, que permitirá desambiguar topónimos tipo GEO/NO-GEO.

5. Compartir la estructura ontológica con la comunidad científica.

## 1.3. Justificación

La necesidad de gestionar información ha sido uno de los factores clave en la consolidación de la informática como una ingeniería imprescindible para el desarrollo de la sociedad. A lo largo de los años se han propuesto gran cantidad de métodos para la desambiguación de topónimos con el objetivo de permitir acceso eficiente a enormes bases de documentos.

A menudo, cuando la información es de tipo textual se incluyen referencias geográficas dentro del texto. Por ejemplo, en las noticias de prensa se suele hacer mención del lugar donde sucedió la noticia. El tener en cuenta estas referencias geográficas proporciona un valor añadido a los sistemas de recuperación de información clásicos.

La importancia de las referencias geográficas reside en las características especiales que tienen debido a su naturaleza espacial. En la investigación en Sistemas de Información Geográfica (*Geographic Information Systems, GIS*) [9] se le ha dedicado mucho esfuerzo al estudio de estas características y al desarrollo de sistemas capaces de aprovecharlas. Este campo ha recibido mucha atención en los últimos años debido, en gran parte, a que las recientes mejoras en el hardware han hecho posible que el desarrollo de este tipo de sistemas sea abordable por muchas organizaciones. Además, dos organismos internacionales *ISO* [10] y el *Open Geospatial Consortium* [11], están realizando un importante esfuerzo colaborativo para definir estándares y especificaciones para el desarrollo de sistemas interoperables. Gracias a todas estas iniciativas muchas organizaciones públicas están trabajando en el desarrollo de infraestructuras de datos espaciales [12] que les permitan compartir su información espacial. Este hecho es una justificación principal del por que trabajar en la desambiguación de topónimos.

Recientemente, ha habido un gran interés en el problema de desambiguación de topónimos desde distintas perspectivas como el desarrollo de recursos para la evaluación de los métodos de desambiguación de topónimos [13] y el uso de estos métodos para mejorar la resolución del alcance (*scope)* geográfico en documentos electrónicos [14], por citar algunos.

No sería posible estudiar la ambigüedad de los topónimos sin estudiar también los recursos que se usan, como bases de datos, diccionarios y otros recursos que se usan para encontrar los significados de diferentes de una palabra. Recursos que a lo largo de este trabajo se han ido estudiando y desarrollando.

A través de la investigación de métodos para la desambiguación de topónimos, la elección del algoritmo apropiado para esta tarea ha sido sumamente importante, debido a que para discriminar las referencias a los lugares puede cambiar en función del recurso elegido y de la información que este puede proporcionar para cada topónimo. Por tal motivo como ya se mencionó anteriormente en este trabajo se implementó un método de desambiguación de topónimos capaz de resolver esta problemática.

## 1.4. Preguntas de investigación

A continuación, se presentan algunas preguntas de investigación concernientes a la presente investigación, estas preguntas guían el proceso mediante el cual se definen los objetivos del desarrollo de la tesis.

a. ¿Qué ventajas tiene la incorporación de objetos geográficos naturales y artificiales a la ontología mixta para ser utilizada como repositorio de sentidos en el desarrollo de la metodología de desambiguación de topónimos?

b. ¿Qué papel juegan las relaciones semánticas espaciales en la tarea de desambiguación de topónimos?

c. ¿Cómo el método propuesto para la tarea de desambiguación de topónimos puede aumentar la precisión con respecto a los resultados que reportan los métodos basados en conocimiento que utilizan corpus como contexto?

d. ¿De qué forma ayudará para dar solución a la tarea de desambiguación de topónimos el modelo de clasificación propuesto?

e. ¿Qué tipo de relación semántica espacial tienen los topónimos de la República Mexicana que aparecen juntos en el mismo contexto?

f. ¿Para qué otras tareas puede ser utilizada la ontología desarrollada?

## 1.5. Alcance e interés del proyecto de tesis

Para gestionar toda la información disponible en ese gran almacén de datos en el que se ha convertido la Web, cuyo crecimiento y descentralización va en aumento, se hace imprescindible afrontar el reto de localizar, procesar e integrar toda la información relevante disponible en este contexto. Puesto que la mayoría de la información en la Web ha sido generada sin ningún tipo de control u organización, es necesario el desarrollo de una tecnología que permita a las computadoras procesar dicho contenido desde un punto de vista semántico para su clasificación y posterior uso.

Los sistemas cada vez más demandan servicios que les permitan situar la información recuperada en un mapa, es decir, *geo-referenciar* la información recuperada a una posición geográfica y espacial asociada (latitud, longitud, entre otros). Además, de que está aumentando el interés en consultas que permitan recuperar documentos relevantes no solo para un tema determinado sino también para una zona determinada. El desarrollo de arquitecturas de sistemas, estructuras de indexación y otros componentes que permitan satisfacer estas necesidades es el objetivo principal de un área nueva de investigación denominada Recuperación de Información Geográfica. Por tanto, es importante que los

sistemas informáticos sean capaces de extraer información geográfica contenida en textos electrónicos. La mayor parte de esta información está formada por nombres de lugares, llamados también topónimos (nombre propio de un lugar).

Esta demanda ha producido que investigadores comiencen a prestar atención en GIR, con el objetivo de proponer arquitecturas de sistemas, estructuras de indexación y otros componentes que permitan desarrollar sistemas mediante los cuales los usuarios puedan recuperar documentos relevantes tanto temática como geográficamente en respuesta a consultas de la forma <*tema, localización*>. La consulta "tesis doctorales sobre sistemas de información geográfica publicadas en Guadalajara" es un ejemplo del tipo de consultas que se estudian en este nuevo campo, que es GIR. El lector familiarizado con los sistemas de recuperación de información clásicos sabrá que la relevancia de los documentos en los motores de búsqueda textual se basa en la frecuencia de aparición de las palabras claves buscadas en los textos y, por tanto, si en un documento no aparece explícitamente la palabra Guadalajara su relevancia se verá disminuida con respecto a esta consulta. Esto sucede aunque aparezca la palabra Jalisco (o cualquier zona metropolitana o municipio de Guadalajara) ya que los sistemas de IR tradicionales no están preparados para tener en cuenta las características especiales de la información espacial (por ejemplo, la relación "contenido en" entre Jalisco y Guadalajara). Esto nos lleva a considerar la ambigüedad que en ocasiones se da entre los topónimos.

La ambigüedad de los topónimos constituye un problema importante en la tarea de la GIR, dado que en esta tarea las peticiones de los usuarios están vinculadas geográficamente. Ésta es probablemente un factor muy importante en la incapacidad de los sistemas GIR actuales por conseguir una ventaja a través del procesamiento de las informaciones geográficas [5].

En este trabajo, se desean abordar temas de interés del área, que están relacionados con la desambiguación de topónimos, considerando características especiales que son debido a la naturaleza espacial de la información con la que se trabaja. Se considerará la naturaleza jerárquica del espacio geográfico y las relaciones topológicas entre los objetos espaciales.

La estructura que permite describir de forma adecuada las características del espacio geográfico es la ontología; para esta tarea en específico, se modela una ontología espacial a través de la incorporación de clases de objetos geográficos naturales y artificiales. De este modo, se estará considerando la naturaleza jerárquica del espacio geográfico y las relaciones geográficas entre los objetos.

Hoy en día, las ontologías han incrementado la atención entre diversos grupos de investigación en el área de la ciencia de la información geográfica. En la actualidad, se

argumenta que las ontologías pueden jugar un rol importante para establecer sólidos fundamentos teóricos y soportar numerosas aplicaciones dentro de ésta área.

## 1.6. Organización del documento de tesis

La metodología seguida para el desarrollo de esta tesis contempla principalmente las etapas del estudio del estado del arte y desarrollando nuevas propuestas. Siguiendo esta metodología y teniendo en cuenta el objetivo general de esta tesis mencionado en la sección 1.2, la organización del resto de la memoria se organiza en otros cinco capítulos.

En el Capítulo 2 se presenta un análisis del estado del arte. Se realiza un análisis a partir de la recuperación de información clásica y recuperación de información geográfica, incluyendo la recuperación de información basada en ontologías.

En el Capítulo 3, se establecen los conceptos básicos y generales que serán utilizados a lo largo de la tesis, con el objetivo de dar al lector un panorama general de una de las tareas importantes dentro de la recuperación de información geográfica, en particular la desambiguación de topónimos.

En el Capítulo 4, se describe la metodología propuesta para el desarrollo de la tarea de desambiguación de topónimos Además, en este capítulo se presenta una descripción detallada de los componentes de la metodología y de la interacción necesaria entre los mismos para realizar las principales operaciones soportadas por la metodología.

En el Capítulo 5, se muestran y discuten los resultados experimentales obtenidos al aplicar el método desambiguador propuesto en el Capítulo 4.

Finalmente, en el Capítulo 6 presentamos las conclusiones obtenidas en la realización de esta tesis y las líneas de trabajo futuro que quedan abiertas.

# CAPÍTULO 2.

# ESTADO DEL ARTE

En este capítulo se incluye una descripción breve de los trabajos más relevantes que se han desarrollado de forma interdisciplinaria para recuperación de información geográfica. Se comentan los enfoques, técnicas, procedimientos y sistemas que se han propuesto por diferentes autores.

# Capítulo 2

# Estado del arte

En este capítulo se revisa el estado del arte de la recuperación de información, un campo de investigación que ha estado activo durante las últimas décadas. De manera particular, se proporciona un estudio de los diferentes trabajos realizados en los últimos años en el área específica de recuperación de información geográfica, la cual por ser un área multidisciplinaria y de reciente nacimiento involucra diferentes líneas de investigación y la convierte en tema de frontera en el ámbito de la investigación en Recuperación de Información Geográfica.

En la figura 2.1, se puede observar de manera más gráfica el contenido de este capítulo y las áreas de como está organizada la discusión acerca del estado del arte.

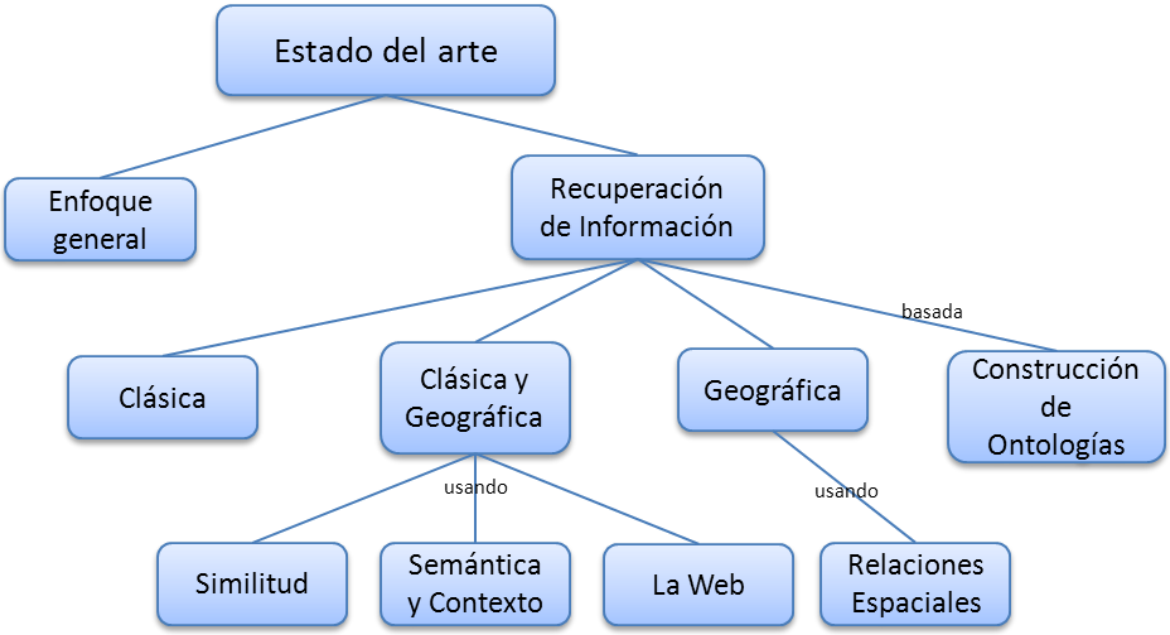


**Fig. 2.1. Mapa conceptual del estado del arte.**

Adicionalmente, esta sección describe los trabajos más relevantes que se han desarrollado de forma interdisciplinaria para recuperar información geográfica. Se comentan y discuten los enfoques, técnicas, procedimientos y sistemas que se han propuesto e implementado. Así como también se señalan las debilidades, aportaciones y áreas de oportunidad actuales en GIR.

## 2.1. Enfoque general

Teniendo como base el análisis realizado sobre los artículos publicados en las áreas de recuperación de información clásica y de recuperación de información geográfica. Se realizarán comentarios acerca de los artículos y trabajos; destacando sus alcances, avances, enfoques, ventajas y desventajas, así como sus aportaciones científicas. En la mayoría de los artículos revisados predominan los enfoques que trabajan con sentidos de las palabras, así como con relaciones semánticas de palabras para datos geográficos en IR y GIR. También, están aquellos cuya meta consiste en resolver la interoperabilidad y la integración de datos desde fuentes heterogéneas.

Otras investigaciones trabajan para mejorar las capacidades (desempeño, relevancia, ponderación) de los sistemas IR. Muchos de estos trabajos se apoyan en estructuras tales como corpus, jerarquías, taxonomías, ontologías y combinaciones entre éstos para realizar IR y GIR. En esta misma línea, destacan los enfoques basados en aspectos cuantitativos y heurísticas.

La mayoría de estas investigaciones se basan en el modelo vectorial, procesos heurísticos y de inteligencia artificial. Aunque, también existen trabajos que usan ontologías como herramientas auxiliares en alguna fase de la recuperación de la información (por ejemplo: en el indexado).

Por otra parte, cabe señalar que los trabajos que manejan ontologías, procesan el conocimiento de éstas, a partir de la implementación y diseño de las mismas. Por ejemplo WordNet[1] es una Ontología cuya implementación está basada en relaciones semánticas de acuerdo a sentidos de palabras. Entonces, el conocimiento se procesa a través de dichas relaciones. También, existen ontologías particulares y generales que son implementadas usando lenguajes como OWL[2]. No obstante, las ontologías se pueden almacenar utilizando estructuras tales como: árboles, redes semánticas, grafos, entre muchas otras. Además, existen enfoques que integran metodologías con soporte de redes semánticas, algoritmos bayesianos, redes neuronales y ontologías. Sin embargo, estos trabajos utilizan en su mayoría datos geocodificados[3] o con un formato definido.

[1] http://wordnet.princeton.edu/
[2] Ontology Web Language. http://www.w3.org/TR/owl-features/
[3] Una geocodificación consiste en asociar un código a un área geográfica o sitio en la tierra (por ejemplo, el código postal).

Además, en estos trabajos se requiere tener un estándar en el formato de datos para que sean fácilmente integrados a otros sistemas, además de disponer de una infraestructura de servicios Web, lo cual es parcialmente alcanzado en Europa o Estados Unidos de América.

Por lo tanto, un enfoque como el anterior, no es factible en países como México, donde los principales proveedores de información geográfica carecen de desarrollos de servicios Web. En general, los sistemas que utilizan servicios Web para tareas de localización se han denominado aplicaciones de búsquedas locales. El mejor ejemplo de este tipo de aplicaciones son las que son ofrecidas por los tres gigantes informáticos: *Google, Yahoo y Microsoft*[4].

Sin embargo, a pesar de que estos servicios han tenido gran aceptación y ofrecen desarrollo continuo, están basados en geocodificación y vinculados a los resultados previamente indexados. Es decir, se basan en un algoritmo basado en geocodificación y no mediante relaciones semánticas, por lo tanto no existe forma de explotar relaciones semánticas en estas aplicaciones. No obstante, algunos de estos servicios de localización son soportados por ontologías de dominio espacial y ontologías espaciales de otras fuentes, donde cierta relevancia es alcanzada, pero con la desventaja de que no consideran el análisis espacial que se le aplica a los datos. Además, de que no consideran la topología de los datos.

Por otro lado, están los denominados servicios RSS[5] (*Really Simple Syndication*), estos servicios aprovechan que esta implementación de XML[6] (*eXtensible Markup Language*) permite difundir información desde diferentes fuentes, pero su funcionamiento, también, está basado en la geocodificación o georeferenciación[7] de los datos. Es decir, si no se tiene el dato geocodificado, entonces este dato no se pude utilizar para tareas de localización usando servicios RSS. En este sentido, la proliferación de estos servicios son el principio del desarrollo de lo que se denomino como la Web Semántica (*Semantic Web*) [15] la cual ha mostrado avances considerables, sin embargo aunque se va consolidando, aún es un reto y no una realidad.

La Web Semántica, también ha sido considerada como una de las soluciones para ciertos problemas en IR, esto debido a que la Web Semántica está orientada hacia la explotación semántica de los datos: en donde se realizan consultas y búsquedas a través del significado de los datos, los cuales aparecen en las consultas (*queries)* y los documentos Web. Cabe destacar, en este punto que bajo este esquema se han realizado muy pocos trabajos enfocados a la semántica de las tareas y procesos en los cuales los datos participan. Es decir, extendiendo el enfoque sintáctico al enfoque semántico, lo cual permite emitir resultados de acuerdo al significado y al contexto de la consulta.

---

Consultar:

[4] {(http://maps.live.com/ ) ( http://local.yahoo.com/ ) ,(http://local.google.com/ )}

[5] Es un formato de datos que es utilizado para redifundir contenidos a suscriptores de un sitio Web.

[6] eXtensible Markup Language, http://www.w3.org/

[7] Un dato georeferenciado es aquél que está asociado con una sistema de coordenadas o proyección en un mapa.

Hoy en día, algunos de los sistemas Web tradicionales ya ofrecen algunas características o servicios basados en la semántica, pero únicamente para datos convencionales (no espaciales). Es decir, que dichos servicios no son factibles de implantarse en un dominio geo-espacial, ya que la naturaleza de los datos espaciales, implica considerar aspectos que en los datos convencionales no son requeridos (por ejemplo: representación, sistema de coordenadas, entre otros). En el caso de GIR y GIS en la Web, también la *GeoSpatial Semantic Web* acuñada así por M. Egenhofer, ha mostrado algunas aportaciones en investigación, pero aún está en proceso de construcción y hoy en día, es considerado uno de los retos más grandes en GIS y GIR, y donde por supuesto se encuentran las áreas de oportunidad para realizar investigación que puede tener alto impacto. Esta tesis trabaja en esta dirección, donde los resultados obtenidos permitirán contribuir en la consolidación de esta nueva línea de investigación.

En otro orden de ideas, en IR y GIR también se trabaja en las fuentes de información (bases de datos, documentos Web y de texto plano, archivos vectoriales). En aspectos tales como la explotación de sus estructuras, propiedades y formatos. Por ejemplo, en el caso de la Web, recuperar información implica considerar la estructura del documento, de acuerdo a formatos tales como: HTML[8], XML, y GML[9] mientras que en el caso de una base de datos se consideran los modelos de datos {*Relacional, Orientado a Objetos, Abstract Data Types (ADT's)*}.

El escenario se complica si se consideran repositorios no estructurados, donde recuperar información, requiere que previamente se hayan organizado, ordenado o indexado los datos. Además de que, los resultados que ofrecen los sistemas actuales (por ejemplo: los buscadores) se presentan con base en la naturaleza textual de la consulta (palabras clave). Lo cual es insuficiente para el dominio geográfico. Lo anterior lo ilustraremos con la siguiente consulta Q= {Hoteles cerca Aeropuerto Benito Juárez} donde la relación "*cerca*" implica una distancia o tiempo. Por lo tanto "*cerca*" debe ser procesada semánticamente por relaciones espaciales, ya que los métodos tradicionales (por ejemplo: sinonimias, clasificaciones, tesauros, etc.) no permiten ofrecer un resultado de acuerdo al espacio geográfico.

Por ejemplo, "*cerca"* puede aparecer en un fragmento como éste: {"*él está cerca de la solución del problema matemático*"}, el cual resulta irrelevante en el dominio geográfico. Sin embargo, también "cerca" puede describirse en fragmentos como: "*Río X está cerca de una población rural*", "*El huracán se aproxima a la población rural a una gran velocidad*", "*El fenómeno meteorológico rodeará la península de Yucatán*" donde cada uno de estos documentos hablan de la relación de "*cerca*", pero la palabra no aparece de forma explícita. En este caso, los métodos (sintácticos) no podrían discernir entre la relevancia de un documento que utiliza el término "*cerca*" en sentido figurado y otro que lo utiliza de acuerdo al dominio geográfico (Esta tarea requeriría procesamiento de lenguaje natural).

---

Consultar:

[8] Hypertext Markup Language, http://www.w3.org/MarkUp/

[9] Geography Markup Language, http://www.opengis.net/gml/

Es por ello que para ofrecer resultados de acuerdo al contexto y dominio geográfico (e.g. Hidrología, Topografía) se requiere considerar otros aspectos como son: las primitivas de representación: puntos (*P*), líneas (*L*) y polígonos (*Pol*) ya que la forma de procesarlas puede filtrar los documentos irrelevantes para una consulta geográfica. Puesto que es diferente medir la "*cercanía"* entre objetos puntos, que entre objetos líneas o polígonos (se requieren tareas adicionales de análisis espacial).

Por ejemplo, para conocer la "*cercanía*" que existe entre un Hotel (*P*) y un Aeropuerto (*P*) se requieren conocer las vías de comunicación (*L*) entre estos. Por lo tanto, se requiere realizar una operación espacial de sobreposición (*Overlay*) entre Hoteles, Ríos y Aeropuertos. Mientras que si lo que se desea saber es el grado de afectación ante el desborde de un Río (*L*) sólo se requiere una operación espacial de buffer[10]. Los procesos anteriores se pueden complicar si, además, se consideran las propiedades geográficas (por ejemplo: los sistemas de coordenadas) donde se requerirían conversiones adicionales, para trabajar con los datos, y posteriormente poder procesarlos bajo un enfoque semántico. Sin embargo, en la actualidad los sistemas de recuperación de información, buscadores y otros sistemas Web no ofrecen soporte para procesar e interpretar búsquedas geográficas basadas en aspectos semánticos.

Esto se debe en gran medida a que los formatos o estructura de los documentos no consideran el comportamiento o el significado de los datos. Además, las representaciones de datos geográficos son en su mayoría de carácter cuantitativo. Lo cual representa un reto importante ya que en las consultas *geo-espaciales* son utilizados términos cualitativos y no cuantitativos (por ejemplo: preposiciones), a continuación se presenta una breve descripción de los trabajos más sobresalientes en cada una de las categorías mencionadas.

[10] Análisis de proximidad o *buffer:* permite medir cercanía, influencia o afectación por un fenómeno climatológico

## 2.2. Recuperación de Información Clásica

En esta sección se comentan cuatro trabajos referidos a la recuperación de información clásica o tradicional. Es decir, La recuperación basada en procesamiento de palabra y aspectos lingüísticos. Estos artículos explican los procesos involucrados para realizar recuperación de información clásica y recuperación de información geográfica, lo cual permite establecer las diferencias existentes estas dos áreas.

### 2.2.1. Kobayashi y Takeda (2000)

En este artículo Kobayashi y Takeda [16] realizan un estudio de las tecnologías y sistemas que actualmente son útiles para la búsqueda y recuperación de la información en la Web. También analizan el desempeño de los mismos de acuerdo a la perspectiva del usuario y al uso de recursos.

Se reporta que el motor de búsqueda es uno de los componentes de mayor investigación para IR en la Web. En donde la mayoría de los algoritmos de recuperación son sintácticos. La investigación está dirigida principalmente a: mejorar la calidad en los resultados (reducir ruido y vínculos rotos) y aumentar la velocidad de recuperación. Considerando, consultas simples, compuestas, híbridas y personalizadas. También detallan como medir el desempeño de un sistema IR utilizando estadísticas y los tres parámetros tradicionales: velocidad, *precisión*, y *recall*. Estos últimos se calculan con las siguientes formulas:

*Precisión:* La proporción de documentos relevantes que fueron recuperados de todos los documentos recuperados, la cual se calcula en la fórmula (1).

$$\text{Precision} = \frac{|\{\text{documentos relevantes} \cap \text{documentos recuperados}\}|}{|\{\text{documentos recuperados}\}|} \qquad (1)$$

*Recall:* La proporción de documentos relevantes que son recuperados, sin considerar todos los documentos relevantes disponibles, la cual se calcula en la fórmula (2).

$$\text{Recall} = \frac{|\{\text{documentos relevantes} \cap \text{documentos recuperados}\}|}{|\{\text{documentos relevantes}\}|} \qquad (2)$$

La Figura 2.2 muestra de forma gráfica los aspectos involucrados en estas fórmulas.

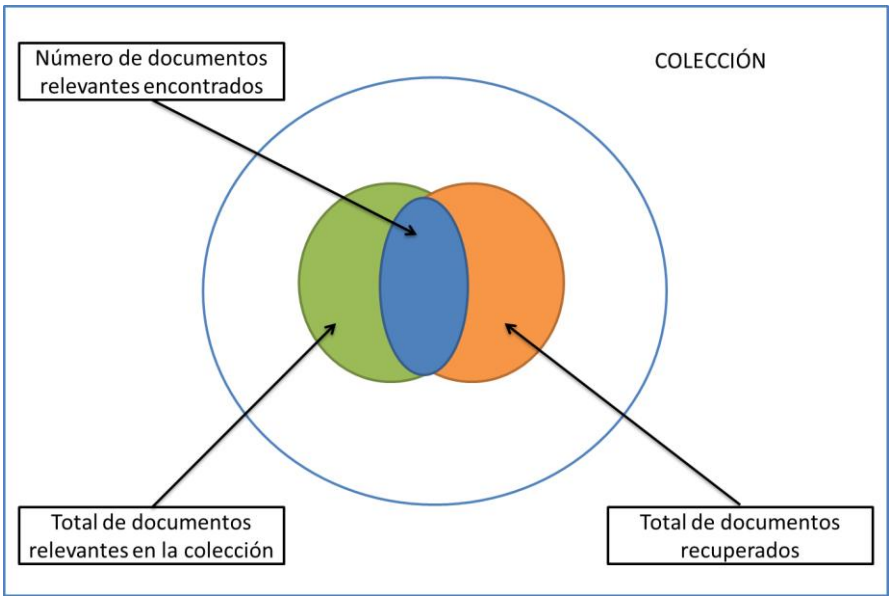


**Fig. 2.2. Representación gráfica de Recall y Precision.**

Se hace hincapié en que algunos enfoques están dirigidos a obtener precisión en los resultados, mientras que otros se centran en conseguir páginas Hub (aquellas que contienen vínculos a páginas relevantes respecto a la información buscada).

Se indican dos de las mayores diferencias entre la IR e IR en la Web:

1) Concurrencia en la búsqueda

2) El número de documentos que pueden ser accedidos y ponderados

Por otra parte, la recuperación en la Web, no está organizada, indexada ni estructurada. Además, una tarea que determine cuales páginas son relevantes para ser indexadas, o cuales son candidatas para asignarles un peso (agruparlas) son procesos de alta exigencia y costo computacional, por lo cual encontrar la solución que represente el menor costo aun no es alcanzada.

### 2.2.2. Braga (2001)

En este trabajo Braga [17] se enfoca en mejorar el proceso de recuperar información específica. En este caso de información relacionada con componentes de software. Para ello realizan un análisis de la búsqueda actual para encontrar componentes de software. Encontrando que las interrelaciones existentes entre componentes de software pueden ser explotadas para mejorar el proceso de recuperación. También, consideran que los componentes están atados a un dominio de aplicación específico para que puedan ser

reutilizados. Por ello, se utiliza una Ontología de dominio para que actúe como una técnica que especifique acuerdos entre componentes de usuarios y proveedores. Utilizando las relaciones semánticas de hiperonimia, hipónimia, y sinónimia.

Resumiendo, el trabajo se enfoca en el problema de interoperabilidad entre repositorios de componentes de software. Usa una arquitectura de capas, donde una de ellas está basada en intermediarios y ontologías que proporcionan el vínculo entre diferentes componentes a los conceptos del dominio. La identificación de componentes relacionados y la organización del dominio, la efectúa cada intermediario, donde cada uno de ellos engloba una ontología de dominio y proporciona el mapeo a sus respectivos repositorios de componentes.

### 2.2.3. Yoel Ledo Mezquita (2006)

Las aportaciones de este artículo [18] consisten en el desarrollo de un nuevo método de desambiguación de sentidos de palabras usando grandes recursos léxicos (diccionarios explicativos, diccionarios de sinónimos, WordNet). Este trabajo está centrado en la resolución de la ambigüedad léxica, la cual aparece cuando las palabras presentan una misma grafía con diferentes significados. A esta tarea se le conoce como Desambiguación del Sentido de las Palabras (*Word Sense Disambiguation, WSD*). La resolución de la ambigüedad de los sentidos de las palabras, es un mecanismo lingüístico para definir el sentido más adecuado de una palabra, según el contexto donde se emplee, que se define en función de los posibles sentidos de las palabras. Por ejemplo, un mecánico de autos busca ¿dónde comprar un gato? y obtiene respuestas sobre "los gatos siameses", "gatos monteses" y otros. Un comerciante de frutas busca "producción de lima" y obtiene respuestas sobre "ciudad de Lima en Perú", "fruta lima", "herramientas para limar metales". Estas imprecisiones son debidas a los distintos sentidos que tienen las palabras.

El método desarrollado contempla la idea del algoritmo de Lesk original, el cual se basa en la búsqueda de intersección de las definiciones de un diccionario explicativo con las definiciones de palabras del contexto en el mismo diccionario. La idea del algoritmo de Lesk simplificado consiste en la búsqueda de las intersecciones de definición de la palabra en cuestión con las palabras del contexto. En resumen en este artículo proponen un método de desambiguación de los sentidos de las palabras que es una combinación de ambos métodos y utiliza adicionalmente la información léxica obtenida de diferentes diccionarios. Además usan la posición relativa de las palabras del contexto para ponderar sus pesos.

### 2.2.4. S.K. Dwivedi y Parul Rastogi (2009)

El objetivo básico de este trabajo [19]) es investigar los aspectos críticos de los diversos enfoques de WSD. En este artículo también se proporciona un detalle de la tasa de éxito de los distintos enfoques para la WSD y también la utilización en los diversos ámbitos de aplicación. Entre los enfoques que contemplan los autores son los denominados métodos basados en conocimiento, donde se hace un breve análisis acerca de Diccionarios de

Lectura Mecánica (*MecMachine Readable Dictionaries, MRDs),* tesauros y recursos léxicos, finalizando con un resumen detallado de estos métodos. También se analizan los métodos supervisados, aquí se contemplan algunos de estos como por ejemplo: clasificador de Naïve Bayes, así como también algunos ejemplos básicos de clasificación (*LazyBoosting Algorithm* y listas de decisión para un clasificador). Y por ultimo analizan los métodos mínimamente supervisados y los no supervisados, como ejemplo de los primeros está el algo ritmo de *bootStrappingApproach* y dentro de los no supervisados se examinan los algoritmos de textos paralelos y los de redes de difusión de activación.

## 2.3. Recuperación de Información Clásica y Geográfica usando similitud

En esta sección se comentan los trabajos que utilizan algún mecanismo de similitud para recuperar y/o ponderar información clásica o geográfica. Estos mecanismos también pueden ser utilizados para ponderar información. En especial existen trabajos enfocados a medir similitud por nombre de palabra, por su semántica lingüística o por sus propiedades geográficas y semántica espacial.

### 2.3.1. Levachkine (2004)

El trabajo publicado en [20] Levachkine describe una forma de representar, analizar, procesar y medir variables cualitativas a través de jerarquías. El artículo en extenso propone este tipo de representación como una nueva estructura de datos, la cual es más simple que las ontologías. De igual forma, las jerarquías son más fáciles de entender y las extensiones a búsquedas y respuestas imperfectas son más sencillas. Adicionalmente, se hace un énfasis primordial a variables cualitativas, las cuales toman valores simbólicos. Estos valores algunas veces pueden acomodarse en capas o niveles de detalle. Por ejemplo, tenemos para la variable *lugar_origen:*

En el nivel 1 toma los siguientes valores: *Europa, África,...,*
En nivel 2 tomas los valores: *Francés, Alemán,…,.*
En el nivel 3 toma los valores: *Californiano, Texano*,…,

En esencia, aquí se considera un dato como una entidad relacional, la cual depende de un contexto particular. Cabe señalar que muchos trabajos sobre procesamiento de datos cualitativos omiten usualmente el contexto del problema; es decir, intentan buscar soluciones generales para todos los contextos de aplicación. Por otro lado, se hace uso de las jerarquías para medir la similitud y disimilitud entre valores cualitativos, intentando conservar el contexto. Para extender la noción de jerarquía, es necesario proporcionar una herramienta adecuada para análisis de datos cualitativos, procesamiento y clasificación; ya

que las jerarquías encapsulan las relaciones (algunas veces ordenadas) entre particiones de un conjunto de datos y por lo tanto mantienen fácilmente el contexto del problema.

Dentro de las principales contribuciones de este trabajo, se tiene que contiene un cálculo de predicados sobre jerarquías, el cual puede ser utilizado para formalizar las consultas soportadas por ontologías. Este último punto es una aportación de esta tesis doctoral. Se propone un método para graduar los errores en consultas, utilizando jerarquías y conjuntos ordenados. Basándose en el hecho de que frecuentemente los valores cualitativos tienen un orden o un nivel jerárquico (muy corto, corto, medio alto, alto).

Por tal motivo, cuando se realizan algunas consultas sobre estas jerarquías pueden presentarse pequeños errores; por lo tanto se proponen una serie de métodos para medir el grado de confusión que puede existir cuando se recuperan datos por medio de consultas. Estos errores pueden suscitarse, cuando se realiza una consulta, ya que ésta puede presentar diversos datos en diferentes niveles de una jerarquía. En esencia para solucionar este tipo de confusiones es necesario contar con un conjunto ordenado de los elementos, esto con el objetivo de evitar en la medida de lo posible el grado de error (*confusión*) en los datos recuperados de una consulta.

En otro enfoque, relacionado con jerarquías en [20], se muestra una forma de cómo introducir conjuntos ordenados o arreglados en jerarquías, las cuales pueden ser utilizadas para diversas tareas:

- Para comparar dos valores, tales como Madrid y Ciudad de México y para medir su confusión. Por ejemplo, para contestar a la consulta ¿Cuál es la capital de España?

- Para comparar la similitud entre dos objetos, utilizando los conceptos de identidad, muy similar, similar, etc., entre diversos objetos.

- Para encontrar la cercanía de un objeto *o* que se pueda adecuar a un predicado.

- Para recuperar objetos que no se adaptan a un predicado dado por un umbral o confusiones acumuladas.

- Para manejar en forma parcial el conocimiento.

En los puntos a destacar de este trabajo encontramos que expresan que las jerarquías realizan buenas aproximaciones para utilizar granularidad de valores cualitativos (en conjuntos ordenados) los cuales proporcionan resultados adecuados en la recuperación de objetos. Adicionalmente, los conjuntos ordenados añaden un refinamiento adicional a la *precisión*, con lo cual la confusión puede ser medida y utilizada.

Además de que esta estructura puede ser empleada como un clasificador de patrones supervisado, utilizando las definiciones de similitud, identidad, etc. Con el objetivo de

medir la cercanía entre dos objetos. Pero, como el trabajo tiene como base el uso de jerarquías, no se describen procesos o metodologías para que la propuesta se pueda aplicar a ontologías. Es por ello que en esta tesis se desarrolla en la metodología un mecanismo para poder utilizar ontologías en lugar de jerarquías.

### 2.3.2. Varelas (2005)

En este trabajo Varelas [21] proponen e implementan un modelo para recuperar información de imágenes y documentos en la Web. Utilizando similitud semántica entre conceptos que no son léxico-gráficamente similares. La determinación de la similitud semántica se realiza con un mapeo entre conceptos y relaciones presentes en la ontología de WordNet.

Se mide similitud con un proceso que detectan similitudes semánticas entre documentos, aunque no contengan términos lexicográficos similares o idénticos. Los resultados ofrecen una mejora mostrada con gráficas de *recall* y *precisión* en relación a otros sistemas de recuperación de imágenes. El modelo se conforma por un módulo de *crawler* (que almacenó 1,5 millones de páginas con imágenes), un módulo de análisis (extrayendo texto y enlaces presentes en cada documento), un módulo de almacenamiento y un módulo de consultas (por palabra clave o texto libre). Dentro de las principales contribuciones que ofrece este trabajo se encuentran:

1) Un *framework* y un sistema implementado para evaluar el desempeño de diversos métodos de similitud semántica usando WordNet.

2) Un modelo de recuperación de información basado en la integración de métodos de similitud semántica.

El modelo propuesto se integró con el modelo vectorial (El modelo vectorial se basa en definir un conjunto de palabras útiles [*keywords* (palabras clave), términos] para con base en este realizar recuperación de documentos. Los documentos se modelan como un vector de términos) en un sistema para recuperar páginas Web e imágenes en la Web. La relevancia de los resultados se midió comparando el modelo propuesto contra el modelo vectorial, a juicio de cinco personas (árbitros). Utilizando veinte consultas que contienen desde uno hasta cuatro términos (no es posible establecer si con un mayor número de términos los resultados serán satisfactorios). Se trabaja con la relación *Is-a* disponible en WordNet, y no se pueden emplear frases o términos compuestos (solo palabras-sustantivos).

### 2.3.3. L. Andrade y M. Silva (2006)

En [22] se describe un operador de similitud geográfica, este operador calcula la relación que existe entre dos lugares geográficos, así como el método para combinar un enfoque geográfico con un enfoque basado en texto para realizar ponderación. La evaluación utiliza la colección de datos del GeoCLEF 2005. Además, definen una estrategia para combinar la ponderación tanto en texto como en geografía. Se utiliza una ontología construida previamente, la cual se explora mediante reglas definidas para asignar la relevancia a los documentos utilizando pesos.

Las pruebas se realizan usando datos del GeoCLEF 2005, mostrando diversas tablas para enfoques de relevancia basados en texto contra el propuesto basado en geografía. Las conclusiones obtenidas son que la ponderación geográfica por similitud y alcance es efectiva únicamente para ciertas consultas geográficas, principalmente las de proximidad. Además el balance óptimo entre ponderación geográfica y textual depende de la consulta. La ponderación textual ofrece buenos resultados sobre la ponderación geográfica, siempre y cuando se procese sobre un número grande de términos geográficos.

## 2.4. Recuperación de Información Clásica y Geográfica usando semántica y contexto

En esta sección se discuten los trabajos que emplean semántica lingüística, espacial, y el contexto para recuperar información textual o geográfica. La revisión de estos trabajos permite tener un panorama claro del papel que actualmente desempeña el procesamiento de la semántica y del contexto. Como resultado de estos trabajos se derivan nuevas líneas de investigación como la del trabajo de la *Web Semántica geoespacial*. Finalmente los trabajos comentados en esta sección permiten visualizar como recuperar información relevante que con otros mecanismos es omitida.

### 2.4.1. Smith & Crane (2001)

En este trabajo de investigación [23] se describe un sistema para desambiguación de topónimos basado en la biblioteca digital Perseo[11]. La categorización de nombres varia significativamente entre los diferentes tipos de documentos, pero la desambiguación de topónimos presenta un alto nivel de *precisión* y *recall*, utilizando un diccionario geográfico de una magnitud mayor que la mayoría de otras aplicaciones.

[11] Disponible en: http://www.perseus.tufts.edu/hopper/

En este método de desambiguación se calcula el centroide geográfico de los candidatos y entonces se remueve todos los candidatos localizados a una distancia mayor al doble de la desviación estándar respecto al centroide.

### 2.4.2. Max Egenhofer (2002)

En este artículo [24] Max Egenhofer describe los elementos necesarios para construir lo que él denomina como la *Web Semántica Geoespacial*, haciendo énfasis en aspectos para la construcción de la semántica formal, así como el desarrollo de múltiples ontologías espaciales y de términos.

Una ontología espacial es aquella que contiene conceptos y relaciones de un dominio geográfico (por ejemplo: Hidrología) mientras que las ontologías basadas en términos se enfocan en la naturaleza del texto (por ejemplo: Wordnet).

Indicando, además que se requieren mecanismos para que la semántica se represente de la tal forma, que sea procesable tanto por personas como por computadoras, así como el procesamiento de consultas espaciales apoyadas por ontologías. Además, se mencionan aspectos tales como la medición de los resultados recuperados basados en la concordancia entre la necesidad de información expresada y la semántica disponible en las fuentes de información y sistemas de búsquedas.

Básicamente, se detalla y describe un nuevo marco de trabajo para la recuperación de información geográfica basada en la semántica de ontologías espaciales y de términos. Adicionalmente, el artículo visualiza y describe la representación de la semántica en diferentes componentes del proceso de recuperación (personas, interfaces, sistemas de búsquedas, y fuentes de información). Se enfatiza en la necesidad de la creación de servicios, en la carencia del procesamiento semántico de datos geográficos, así como también en como el papel de las ontologías puede ayudar a construir el equivalente a la Web Semántica. Finalmente, se menciona que la *Web Semántica geoespacial* permitirá a los usuarios recuperar de forma más precisa los datos que ellos necesitan, basándose en la semántica asociada a dichos datos. Se subrayan diversos retos para lograr esto, en esencia son áreas de oportunidad en GIR.

### 2.4.3. Guha (2003)

En este artículo Guha [26] presentan una aplicación de búsqueda semántica diseñada para mejorar los resultados obtenidos por la búsqueda tradicional en la Web. El sistema considera la notación del *query* de búsqueda e incrementan los resultados de la búsqueda usando datos relevantes obtenidos de diferentes fuentes. Subrayando que la explotación de la semántica de cada término de una consulta mejorara la recuperación de información.

Se enfatiza el hecho de que los resultados relevantes aumentan para una búsqueda que considera la semántica. Cabe destacar que la dirección de este trabajo está enfocada en el componente de búsqueda basada en texto y se propone un sistema hibrido que procese palabras clave y la semántica de una notación particular para una consulta específica. Finalmente, este trabajo requiere ser adaptado para ser empleado en el dominio geográfico.

### 2.4.4. Rocha (2004)

En este artículo [27] se presenta una arquitectura de búsqueda que combina técnicas de búsqueda clásicas con técnicas de *"spread activation"* aplicadas a un modelo semántico para un dominio específico.

La técnica de "*spread activation*" permite, principalmente, encontrar conceptos relacionados en la ontología, logrando esto a través de un conjunto inicial de conceptos (proporcionado) y sus correspondientes valores de activación iniciales. Estos valores iníciales son obtenidos de los resultados de búsqueda clásica aplicados a los datos, los cuales son asociados con los conceptos presentes en la ontología. Para la realización de pruebas se implementaron dos escenarios de propagación basado en enfoques simbólicos y sub-simbólicos, denominada activación híbrida. El segundo escenario se realizó con un enfoque único, que puede ser simbólico o sub-simbólico. En este último los resultados se indican como positivos. Con base en los resultados se concluyó que la activación híbrida propuesta, alcanzó mejores resultados que el resto de los enfoques.

El sistema usa en enfoque de expansión de consulta donde el primer conjunto de expansiones busca mejorar la funcionalidad existente a través del uso de diferentes pesos. Mientras que el segundo conjunto de expansiones debe agregar nuevas características tales como: incorporación en una categoría basada en un peso, mapeo de pesos de acuerdo al contexto, retroalimentación por relevancia, etc.

Finalmente, una característica interesante de la técnica de "*spread activation*" es que es posible proponer un conjunto de conceptos que se estimen estén fuertemente conectados para un concepto dado, aunque no exista una relación explícita entre los conceptos almacenados en la base de conocimiento. Pero, no se ofrecen mayores detalles acerca del cual sería el criterio que permitiría "estimar" un fuerte acoplamiento o conexión entre ciertos conceptos.

### 2.4.5. Reiner Kraft (2005)

En este artículo, Reiner [28] proporcionan un panorama general del sistema *Y!Q Contextual Search*, subrayando las técnicas utilizadas para capturar el contexto de una búsqueda (el contexto es un conjunto de palabras relacionadas con los términos que conforman la consulta) y con ello mejorar la recuperación de información.

La búsqueda contextual intenta capturar de mejor manera las necesidades de información del usuario. A través de expandir la consulta con información contextual. El artículo se enfoca en como capturar el contexto de búsqueda con alta calidad, y cómo utilizar este contexto para mejorar la relevancia de los resultados. Para el primer problema, Y!Q presenta un elemento gráfico que captura el contexto de búsqueda y proporciona acceso a su funcionalidad en el punto de interés de la consulta. *Y!Q* utiliza una red semántica para analizar el contexto de búsqueda, resolviendo posibles ambigüedades en los términos, y generando un compendio contextual que consiste de sus conceptos clave.

Este compendio, es enviado a un planificador de consultas y marco de reescritura para expandir la consulta del usuario con términos de contexto relevantes, y así mejorar la relevancia global de la búsqueda. Los resultados de *Y!Q* se comparan con los resultados del buscador Yahoo! En donde *Y!Q* obtiene una mayor relevancia. En particular, indican que para consultas ambiguas el contexto ayudó a proporcionar mayor relevancia y dirigir resultados de acuerdo al contexto.

### 2.4.6. Clough (2005)

En [29] se propone una heurística basada en el cálculo de la puntuación de solapamiento entre el contexto y la ruta de acceso al referente jerárquico (es decir, el numero de topónimos en común). La puntuación más alta, significa que es más probable que el referente sea correcto. La aproximación final, filtra candidatos a lugares mediante reglas de contexto para eliminar palabras vacías, las referencias a personas y organizaciones y enlaces a los correos electrónicos (URL). Por ejemplo Sr. Sheffield se filtra como no-geográfico y se referencia es utilizando la siguiente regla de contexto: < *title* > < *location* > *null*, donde titulo y localización son marcadores de posición para entradas en una lista del gazetter.

### 2.4.7. Leidner (2008)

Este artículo [13] Leidner, se concentra en nombres geográficos de lugares populares, donde se define la tarea de Resolución Automática de Topónimos (*Toponym Resolution*, *TR*) como el cálculo del mapeo de las apariciones de los nombres de los lugares que se encuentran en un texto hacia su representación semántica extensional de la ubicación a la que se refiere, así como su *footprint* de latitud y longitud. La tarea del mapeado de nombres hacia sus localizaciones es difícil debido a las bases de datos insuficientes y ruidosas, y con un alto grado de ambigüedad (geo/no-geo ambigüedad), y el mapeo entre nombres y localizaciones es ambiguo, por ejemplo London puede referirse a la capital de UK o a la de London, Ontario, Canadá o algunos otros London en la tierra. El objetivo principal es investigar como referenciar nombres espaciales ambiguos de entidades que pueden estar bien fundamentadas, o resueltas, con respecto a un modelo robusto coordinado extensional de dominio abierto de noticias.

### 2.4.8. Buscaldi (2010)

El objetivo de este trabajo [5] es estudiar la ambigüedad de los topónimos y los efectos de su resolución sobre aplicaciones como la GIR, la búsqueda de respuestas y la recuperación de información en la web. En este trabajo se desarrolla un método basado en densidad conceptual y otro basado en la distancia media desde centroides en mapas. La densidad conceptual es una medida de correlación entre el sentido de la palabra (camino jerárquico de topónimos candidatos en WordNet) y su contexto (GeoSemCor). Se calcula en las sub-jerarquías de Word- Net, determinada por la relación de hiperónimos.

## 2.5. Recuperación de Información Clásica y Geográfica en la Web

En esta sección se comentan los trabajos que recuperan información geográfica y clásica desde la Web. En donde se proponen diversos métodos para resolver problemas de ubicación, de indexación y de recuperación basada en aspectos geográficos.

### 2.5.1. Hawking (2001)

En este trabajo Hawking [30] realizan una evaluación de veinte máquinas de búsqueda con el objetivo de evaluar la calidad de cada una de éstas. La evaluación se realizó con base en un conjunto de cincuenta y cuatro *queries* tomados de los archivos *log* de ciertos buscadores Web.

Además, para la evaluación se tomaron en cuenta sólo los primeros siete resultados de cada máquina de búsqueda. Particularmente, el estudio presentado y sus predecesores manejan consultas que se asume son obtenidas de la necesidad de encontrar una selección de documentos relevantes para un tema especifico. Enfatizan en que la meta es encontrar nuevas técnicas para implementarse en los buscadores actuales, así como mejorar sistemas que usan metodologías bien conocidas de IR. Las medidas usadas para la evaluación de los motores de búsqueda se relacionan principalmente con el desempeño del *crawler*, y se concluye que ninguna de las máquinas de búsqueda más populares explora más del 16% de su estimado total (aproximadamente 800 millones de páginas Web indexadas) para satisfacer una consulta.

Los metabuscadores no indexan documentos por sí mismos sino que reenvían las consultas a las máquinas de búsqueda más populares y conforman una lista de resultados. Donde el criterio de comparación para las máquinas de búsqueda consideradas usa un rango de medidas obtenidas a través de juicios de relevancia binaria. Lo cual en IR es una desventaja ya que no hay forma de aproximar, es decir en un enfoque booleano, solo se puede

categorizar en dos valores: bueno o malo. Básicamente lo que se hizo fue medir precisión de una lista de resultados obtenida por once máquinas de búsqueda, donde destacan: *Yahoo!, Google y Microsoft*.

### 2.5.2. Jones (2002)

En este artículo Jones [25] reportan algunos métodos para extraer información geográfica desde páginas Web como parte del proyecto SPIRIT (*Spatially-Aware Information Retrieval on the Internet*). Este proyecto consiste en la recuperación de la información espacial a través de ontologías geográficas, donde incluso ontologías como *WordNet* (basada en sentidos de palabras en inglés) han sido ampliamente explotadas para integrarse a este proyecto.

El objetivo del proyecto es enriquecer los sitios Web usando conceptos espaciales y construyendo conjuntos de datos espaciales que estén visibles y disponibles en Internet. Así como realizar recuperación de información considerando criterios espaciales y geográficos. El proyecto resultó innovador, ya que a través de conceptos se recupera la información. Sin embargo, en el marco general de estos conceptos sólo se consideran atributos descriptivos definidos para cada objeto geográfico tales como: nombres, direcciones, código postal, números telefónicos, etc. Y hasta el momento, no reportan atributos espaciales o la consideración de relaciones espaciales (enfocadas a topología) entre los objetos geográficos. Los esfuerzos hasta el día de hoy son enfocados a criterios geográficos y de localización.

En esencia, este trabajo describe una herramienta espacial para extraer *metadatos*, y un programa de geocodificación que asigna coordenadas espaciales a cada una de las localidades extraídas por algún método especifico. Este último aspecto permite ubicarlo en la categoría de sistemas de georeferenciación. El proyecto descrito es un sistema GIR que se apoya en ontologías léxicas y jerárquicas, tesauros, ontologías espaciales y dominios geográficos. Muy al estilo de WordNet. En fechas recientes han publicado trabajos enfocados con el manejo de relaciones espaciales y algunos enfoques semánticos, pero el procesamiento de relaciones se lleva a cabo de forma aislada y sin considerar tareas de análisis espacial.

### 2.5.3. Zhou (2005)

En este trabajo Zhou [31] se enfocan en las búsquedas en la Web considerando su ubicación y no la descripción de su ubicación, es decir, el lugar donde residen o están alojadas las páginas Web. Por ejemplo: la página de la embajada mexicana, en un servidor de Canadá. El objetivo consiste en encontrar información contenida en páginas Web cuyos temas están relacionados con un lugar o región.

Justifican su investigación en el hecho de que los motores de búsqueda convencionales trabajan con indexado orientado a conjuntos, mientras que la información relacionada con una ubicación es de dos dimensiones en el espacio Euclidiano. Por lo tanto, se enfocan en la representación eficiente de los atributos de ubicación presentes en documentos Web. Además de establecer una forma de combinar ambos tipos de indexado (textual y de localización). La propuesta consiste en una estructura de indexado híbrida la cual integra archivos invertidos y árboles R*[12].

Donde un archivo invertido (índice invertido) es una estructura de índice que almacena el mapeo de palabras a sus ubicaciones en un documento o conjunto de documentos lo cual permite búsqueda de texto libre. Por ejemplo, dados los textos $T_0$ = "esto es que esto es", $T_1$ = "que es esto" y $T_2$ = "esto es una manzana", el archivo invertido se presenta en la figura 2.3.

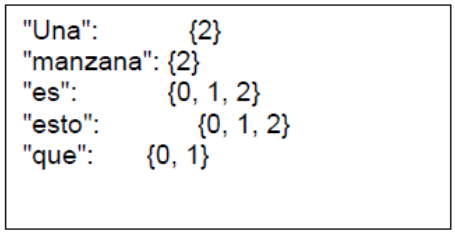


**Fig. 2.3. Estructura de un archivo invertido [31].**

Tal y como se aprecia en la figura 2.3 se almacena el número de texto (documento) en el cual una palabra o término aparece. Entonces en este trabajo se integra archivos invertidos y árboles R* para gestionar las consultas relacionadas con ubicación y consultas textuales. Además, se indican que se estudiaron las posibles combinaciones resultando en tres esquemas:

1) Archivos invertidos e índices de árbol R*

2) Primero archivo invertido y después árboles R*

3) Primero árboles R* y después archivos invertidos.

De manera adicional, se propone un esquema para validar el desempeño de la estructura de indexación propuesta, la cual consiste en un motor de búsqueda Web conformado a su vez en cuatro módulos:

[12]Los árboles-R* o R*-árboles son estructuras de datos de tipo árbol similares a los árboles-B, con la diferencia de que se utilizan para métodos de acceso espacial, es decir, para indexar información multidimensional. Por ejemplo, las coordenadas (x, y) de un lugar geográfico.

1) Un extractor el cual detecta alcances geográficos de las páginas Web y representa el ámbito geográfico con base en coordenadas geográficas.

2) Un indexador el cual construye las estructuras de indexado híbridas para integrar información de ubicación y textual.

3) Un ponderador que trabaja con base en relevancia geográfica y no geográfica.

4) Una interfaz que permite visualizar la relevancia de los resultados geográficos y no geográficos.

En esencia, el trabajo se centra en la hipótesis que si se busca cierta información desde la ciudad *X*, es más probable que el usuario se interese solo en los sitios de la ciudad *X*, por lo cual se descartan las ciudades *Y* ó *Z*. La desventaja aquí, es que se requiere tener la información de la red y además procesar la información de archivos log, para determinar desde donde está accediendo a un cierto sitio un usuario.

### 2.5.4. T.Delboni (2005)

En este documento [32], se explora el uso de las expresiones en lenguaje natural para realizar búsquedas geográficas en la Web, en particular, enfatizan en el hecho de que no se utilizan datos geocodificados. Las consultas están referidas a la localización de sitios usando términos utilizados para referirse a lugares y su ubicación sin mencionar coordenadas, por ejemplo: la consulta "Aeropuertos cercanos a Brasil". Estas expresiones denotan la posición de un sujeto con respecto a un punto de interés (*landmark*), por ejemplo: el edificio de gobierno es un punto de interés o referencia para ubicar otro sitio cercano a éste.

La justificación para procesar este punto de interés como el componente que permita procesar la semántica de los términos relacionados con cercanía, reside en la aseveración de que es una fuente valiosa del contexto geográfico que está incrustado en muchos documentos Web. El enfoque propuesto, tiene la finalidad de guiar hacia una técnica basada en la expansión de consultas, la cual utiliza una máquina de búsqueda que trabaja bajo un mecanismo de *keyword-matching*. Se presenta una interfaz del sistema prototipo para sitios en Brasil. En resumen, en este trabajo se explora el uso de las expresiones que de forma común una persona utiliza para ubicar un lugar, en particular usando puntos de referencia, enfatizan en el hecho de que no se utilizan datos geocodificados. Básicamente lo que se hace es asociar a términos de localización un valor numérico. Por ejemplo: si se habla de cercanía se asigna mediante un proceso, un valor a este término para poder satisfacer el significado de cercanía.

El trabajo pertenece al enfoque de expansión de consulta. Una de las desventajas reside en saber cuántos términos son necesarios en la expansión de consulta, y por otro lado, que existe dependencia de un motor de búsqueda basado en palaba clave. Así que el trabajo

ofrece mejoras en la relevancia de resultados, pero estas mejoras son basadas en procesos estadísticos y aspectos cuantitativos.

### 2.5.5. Vaid (2005)

El presente artículo [33] presenta tres métodos para realizar indexado tanto espacial como textual. Está orientado hacia máquinas Web de búsqueda, considerando que éstas tratan los términos geográficos en la misma forma que otros términos, lo cual genera que los resultados no sean los requeridos por el usuario. Entonces bajo esta premisa se propone asociar métodos de indexado espacial con los de indexado textual, explotando los procedimientos de *geo-tagging* para categorizar documentos con respecto al espacio geográfico. Sin embargo, los esquemas son comparados experimentalmente con máquinas de búsqueda convencionales (que usan indexado textual) para mostrar su desempeño y velocidad con respecto a aquellas que utilizan únicamente indexado textual. Además, las pruebas no fueron evaluadas mediante criterios espaciales, sino que se usaron algunos buscadores que explotan el texto, por lo tanto es claro que la relevancia de los resultados mejora, pero es necesario medir esta relevancia con otros sistemas que si consideren aspectos geográficos y no solo mediante los aspectos del texto.

### 2.5.6. D. Santos y M. Chaves (2006)

En este trabajo [34] se discuten los métodos usados en GIR para realizar geo-indexación con páginas Web, utilizando el método de *Geo-scoping* (*Grounding*). En donde demuestran que el método no ofrece los mejores resultados cuando la página Web de una ciudad *X* describe una ciudad *Z*, pero la página Web no está alojada en la ciudad *Z*. Por ejemplo, un sitio mexicano cuya página habla de Suiza, y por otro lado, la página de una tienda situada en Suiza y que vende zapatos en ese país. Se discute el problema de lo que significa "lugar" en lenguaje natural.

Los autores investigan el rol de "lugar" en lenguaje natural integrando ocho aspectos, que han sido tratados de forma independiente en otros trabajos. Se apoyan en una máquina de búsqueda existente, una ontología de lugares, un sistema de reconocimiento de nombres de entidades (NER) y una colección de datos en portugués. Las pruebas que realizan son reportadas de forma empírica, considerando: distribución de distritos y división política de Portugal, una geo-Ontología basada en mapas, y conceptos geográficos de textos portugueses. Estos últimos son dependientes de la cultura (serían diferentes para otros países). Los resultados indican cuantos documentos mencionan sitios geográficos por nombre, cuantos se repiten, tipo de sitios se refiere (ciudades, villas, etc.) y el porcentaje de ambigüedad para los nombres de lugares y de organizaciones. Finalmente, se concluye que los diferentes roles para nombres de lugares en lenguaje natural, son dependientes del lenguaje y de la cultura. Además de que son expresados de forma imprecisa, y son dependientes del contexto.

### 2.5.7. Jens Graupmann and Ralf Schenkel (2006)

En este trabajo [35] describen un motor de búsqueda para consultas geográficas en la Web. El cual incluye una interfaz visual de los resultados discretos o expandidos para una consulta específica. Contrario a otros enfoques, el artículo no asigna *footprints* a los documentos, pero si considera el contexto de la información geográfica de forma general, con lo que se permite evaluar consultas de granularidad fina, esto al nivel de fragmentos de documentos. No permite hacerlo con documentos completos.

El trabajo se enfoca en consultas geográficas la cuales combinen restricciones sobre el contenido de una página Web con al menos una restricción geográfica, por ejemplo usando rangos ("*between Paris and Nancy*") e imprecisión ("*near London*"). El motor de búsqueda presentado es una extensión a un motor de búsqueda ya existente (*SphereSearch Engine*, *SSE*) donde SSE integra técnicas de IR y de extracción de información (IE). La propuesta de solución consiste en realizar anotaciones de información geográfica dentro de una página y agregándola en el contexto de un contenido que coincidió (match) al explorarlo, donde posiblemente se incluye conocimiento externo como una jerarquía de ubicaciones.

## 2.6. Recuperación de Información Geográfica usando relaciones espaciales

En esta sección se describen los trabajos relacionados con recuperar información geográfica desde una fuente de datos vectorial. Asimismo se describen enfoques que explotan las relaciones espaciales entre objetos geográficos. Cabe señalar que son muy pocos los trabajos enfocados a procesar las relaciones espaciales. Siendo entonces esta una área de oportunidad importante en GIR.

### 2.6.1. Walker (2005)

En este trabajo Walker [36], proponen el uso de algoritmos de aprendizaje bayesianos[13] como una alternativa para mejorar la recuperación de información geográfica. Cabe destacar que este artículo utiliza las relaciones espaciales como un factor de importancia para recuperar información geográfica. La propuesta considera la tarea de análisis espacial más común en GIS (la sobreposición). La cual se realiza sobre un conjunto de temas (capas de datos) también conocidos como "*workspaces*" o "*projects*" en software comercial. La propuesta está dirigida a automatizar tareas de análisis espacial vertical, es decir aquellas que se aplican sobre un conjunto de capas de datos.

[13] Trabajan con base en probabilidades que son establecidas en cada nodo que conforma a una red.

En particular, es una alternativa de solución al problema de decidir cuales temas (capas) deben ser incluidos en un mapa (*workspace*) para realizar una tarea de análisis espacial específica. Por ello, enfatizan que las técnicas actuales de GIR, recuperan las capas de datos aisladamente y se agregan manualmente a un mapa. Pero, resulta más significativo analizarlas en conjunto. Sin embargo, para definir las capas de datos que conformarán un mapa es necesario recuperar la información más relevante de entre todas las capas disponibles, y es en este punto donde tiene origen este trabajo.

No obstante, en la práctica los temas de un mapa no son analizados de forma individual, entonces su propuesta consiste en analizar el conjunto de temas que conforman un mapa previamente creado. El propósito de dicho análisis es extraer las relaciones espaciales presentes en estas capas de datos, para después explotarlas y a partir de ellas inferir cuales capas de datos son las más ideales para la construcción de un mapa. La justificación para argumentar que la extracción de las relaciones espaciales de los temas que conforman un mapa permitirá automatizar dicho proceso, se basa en que el analista GIS construye mapas de acuerdo al tipo de análisis que realizará. Entonces para tareas de análisis diferentes, los mapas contendrán algunos de los temas usados en otra tarea, es decir coincidirán en las capas usadas. Además, hacen referencia a que las redes bayesianas han sido adoptadas para asignar relevancia a un tema, de forma tal que el enfoque se puede usar para automatizar la creación de mapas.

En esencia, el trabajo propone tres algoritmos espaciales de aprendizaje de redes bayesianas, los cuales incorporan las relaciones espaciales presentes en los temas, en el proceso de aprendizaje. Las redes bayesianas resultantes fueron cargadas en una máquina de inferencia que se utilizó para recuperar todos los temas relevantes, de un conjunto de prueba, para una consulta de usuario. El rendimiento de los algoritmos de aprendizaje espacial bayesiano se evaluó y comparó contra los algoritmos convencionales (aquellos que no manejan aspectos espaciales). Una aportación interesante en el trabajo es que la recuperación de la información, se realiza explotando y trabajando con datos vectoriales y sobre la tarea de análisis espacial de sobreposición, y es el único que ha trabajado con este enfoque.

## 2.7. Recuperación de Información apoyada en la construcción de Ontologías

Esta sección comenta los trabajos que han utilizado alguna metodología o mecanismo para construir o procesar ontologías. Las fuentes de información utilizadas describen los datos principalmente mediante palabras.

### 2.7.1. Sang Ok (2003)

En este artículo Sang Ok [37], proponen un método semi-automático para la construcción de una ontología usando agrupaciones de palabras (*hub words)*. El enfoque tiene como hipótesis la definición de ciertas palabras (*hub words*) las cuales están relacionadas con un conjunto de otras palabras. Estas *hub words* son determinadas por la frecuencia de aparición de los términos en un documento. Un detalle interesante es que se propone un proceso de construcción para una ontología usando *hub words* y un método automático para la extensión de la ontología (agregando relaciones). Además de que resaltan que la ontología propuesta se puede usar como un archivo de indexado en IR, justificando esta afirmación en el hecho de que una ontología puede ofrecer mayor información semántica que los archivos índices.

El artículo describe la construcción de la ontología y su extensión, así como el proceso de recuperación de información. La construcción de la ontología se realiza en tres pasos:

1) Encontrar las palabras con mayor índice de frecuencia en una colección de texto (sustantivos)

2) Encontrar manualmente la ontología donde los nodos principales son las *hub words*

3) Extender automáticamente la ontología (agregar las relaciones)

Este proceso se realiza agregando las palabras que tengan relaciones (acorde al dominio) con las *hub words* seleccionadas.

Por otra parte, el proceso de extracción de *hub words* se basa en técnicas conocidas de IR, sin embargo la aportación consiste en considerar el dominio del problema, ya que las *hub words* seleccionadas para un dominio económico serán diferentes para un dominio como la medicina. Por ejemplo, en un dominio económico la palabra "*company*" tiene muchas relaciones con palabras tales como: "*share*", "*interests*", "*stock*", "*trade*", "*lawyer*",…entre otras, pero en el dominio de la medicina tiene pocas relaciones.

El proceso de extensión de la ontología consiste en agregar relaciones, siguiendo los mismos pasos que en la construcción de la ontología, pero considerando los verbos que

aparecen entre *hub words* y las no *hub words*. El primer paso extrae los sustantivos vecinos de las *hub words* (se aplican reglas de análisis de enunciados) y el segundo paso establece dicha relación. Por ejemplo, si el verbo "*belong*" o "*include*" aparece entre dos sustantivos, entonces la relación "*belong to*" se considera como candidata a relación entre las *hub words* que contienen dichos sustantivos.

Se utiliza RDF para realizar las consultas en la ontología y se presentan gráficas que muestran los resultados obtenidos. Un detalle de este trabajo es que no se explica cómo se establecieron las reglas de extracción para las ontologías (las cuales serían muy valiosas). Además de que las técnicas usadas son estrictamente de análisis de texto.

### 2.7.2. Maria A. Leite & Ivan L. M. Ricarte (2008)

El objetivo de este artículo [38] es explorar un marco de trabajo para codificar una base de conocimiento geográfico, compuesto por múltiples ontologías relacionadas, cuyas relaciones se expresan como difusas. Cada ontología representa un área distinta de conocimiento relacionada con referencias geográficas. Esta organización de conocimiento se utiliza en un método difuso para expandir la consulta inicial del usuario. Cada ontología puede ser representada independientemente así como también sus relaciones

### 2.7.3. Diego Seco Naveiras (2009)

En este trabajo [39] se abordan varios temas de interés en el área de GIR. En primer lugar, las estructuras de indexación que permiten recuperar documentos empleando tanto su ámbito textual como su ámbito espacial, no tienen en cuenta la naturaleza jerárquica del espacio geográfico ni las relaciones topológicas entre los objetos espaciales que indexan. Por tanto, su primer objetivo fue desarrollar una estructura que solucione los problemas debidos a limitaciones. Se desarrolló un prototipo de sistema, basado en una arquitectura genérica, modular y extensible.

En este trabajo se formaliza una ontología del espacio geográfico en la que se basa la estructura de indexación, la ontología proporciona un vocabulario de clases y relaciones para describir un ámbito determinado para el caso el espacio geográfico. El objetivo no es definir una ontología que describa todo el espacio geográfico sino sólo para el ámbito determinado en el que se vaya a utilizar el sistema. Aunque en el prototipo se defina una ontología y una estructura de indexación para un domino general donde se considere que los niveles que se deben tener en cuenta son los continentes, los países, las regiones y el resto se consideran parte de la categoría de lugares poblados.

La ontología describe ocho clases de interés: *SpatialThing, GeographicalThing, GeographicalRegion, eopoliticalEntity, PopulatedPlace*, *Region, Country* y *Continent.* Además, existen relaciones jerárquicas entre *SpatialThing, GeographicalThing, GeographicalRegion* y *GeopoliticalEntity* ya que: *GeopoliticalEntity* es subclase de

*GeographicalRegion*, *GeographicalRegion* es subclase de *GeographicalThing* y *GeographicalThing* es subclase de *SpatialThing*. Es decir, estas cuatro clases están organizadas en una jerarquía de especialización superclase-subclase, también conocida como taxonomía.

### 2.7.4. López (2012)

En esta investigación [40] se toma en cuenta la problemática de GIR, que sugiere el desarrollo de técnicas de desambiguación de topónimos basadas en ontologías para tratar la ambigüedad en consultas de GIR. En particular se construyó una ontología de dominio mixta, a manera de organizar los cerca de 500 millones de hispanohablantes, con el fin de desarrollar un desambiguador de topónimos para el lenguaje Español. El desambiguador propuesto se basa en proximidad geográfica entre topónimos del mismo contexto, usando relaciones jerárquicas que proporciona la ontología, y al mismo tiempo en una ponderación jerárquica ontológica, complementada con la distancia de Haversine.

# CAPÍTULO 3.
# DESAMBIGUACIÓN DE TOPÓNIMOS

En este capítulo se describe el marco teórico del trabajo de investigación de esta tesis, la desambiguación de topónimos, se concentra en proporcionar las definiciones y teorías apropiadas que se utilizan para el desarrollo del presente trabajo.

# Capítulo 3

# Desambiguación de Topónimos

La desambiguación de topónimos constituye una de las tareas importantes dentro de la recuperación de información geográfica. Recientemente, ha habido un gran interés en este problema desde distintas perspectivas como el desarrollo de recursos para la evaluación de los métodos de desambiguación de topónimos [13] y el uso de estos métodos para mejorar la resolución del alcance (*scope*) geográfico en documentos electrónicos [14], por citar algunos. No sería posible estudiar la ambigüedad de los topónimos sin estudiar también los recursos que se usan, como las bases de datos, diccionarios y otros útiles en el proceso para encontrar los diferentes significados de una palabra. A lo largo de la investigación de métodos para la desambiguación de topónimos, la elección del algoritmo apropiado para esta tarea ha sido sumamente importante. El recurso léxico elegido influye en gran medida en la discriminación de las referencias a los lugares.

Considerando que los métodos para la desambiguación de topónimos son muy diferentes pero con factores en común, la mayoría están influenciados por dos fases principales mencionadas a continuación:

1) **Extraer los referentes candidatos**. Todos los posibles referentes al lugar son extraídos de un recurso de conocimiento geográfico (repositorios de sentidos: gazzetteer, ontología, etc.).

2) **Escoger el referente correcto**. Se aplican una serie de heurísticas para determinar de entre todos los posibles candidatos, aquel que más probablemente tenga el significado correcto de acuerdo al contexto (corpora textual) y a los recursos como fuente de evidencia.

En base a lo anterior, en este capítulo se incluye una descripción acerca de la desambiguación de topónimos para tener un panorama más general sobre esta tarea.

## 3.1. Midiendo la ambigüedad de topónimos

¿Qué tan grande es el problema de la ambigüedad de topónimos? En cuanto a la ambigüedad de otros tipos de palabras en las lenguas naturales, la ambigüedad de topónimos está estrechamente relacionada con el uso que la gente hace de ellos, afirma Buscaldi [5]. Por ejemplo: un mesero puede ignorar que "copa" no es sólo un vaso con pie para beber, si no también la parte más alta de un árbol. De la misma manera, muchas personas ignoran que "Guadalajara" es una ciudad de la República Mexicana, pero también una ciudad o municipio de España, que en algunos casos puede conducir a errores.

Los diccionarios[14] pueden ser utilizados como una referencia para los sentidos que pueden ser asignados a una palabra, en este caso a un topónimo. Un problema con los topónimos es que la granularidad[15] de los diccionarios geográficos (*gazetteers*) puede variar considerablemente de un recurso a otro, el resultado de la ambigüedad de un topónimo puede no ser el mismo con la utilización de diferentes gazetteers. Por ejemplo, Smith y Mann [41] estudiaron la ambigüedad de los topónimos a nivel continental con Getty TGN, logrando casi que el 60% de los nombres utilizados en el Norte y Central de América fueron ambiguos (es decir, para cada topónimo existen al menos 2 lugares con el mismo nombre). Sin embargo, si la ambigüedad topónimo se calcula sobre nombres geográficos, estos valores cambiar significativamente.

En la tabla 3.1 se muestran los topónimos más ambiguos de acuerdo a GeoNames, GeoPlanet y Wornet [16], respectivamente. De acuerdo a este cuadro se puede apreciar el nivel de detalle de los distintos recursos, ya que hay 1,536 lugares nombrados "San Antonio" en GeoNames, casi 7 veces más que en GeoPlanet, mientras que en Wordnet el topónimo más ambiguo sólo tiene 5 posibles referentes. Esto da una idea más clara de que dependiendo el recurso que se utilice es el resultado que se tendrá a la hora de desambiguar topónimos. De esta tabla también se puede observar que los topónimos como San Antonio, San José, Santa Rosa, San Francisco, Benito Juárez, Santa Cruz, Guadalupe, San Isidro, y Victoria son topónimos ambiguos de la República Mexicana.

Continuando con esta tabla (tabla 3.1) se puede ver que "San Francisco" es uno de los topónimos más ambiguos tanto para GeoNames como para GeoPlanet. Sin embargo, ¿es posible afirmar que "San Francisco" es un topónimo muy ambiguo? La mayoría de las personas en el mundo probablemente sabe sólo el "San Francisco" en California y no el "San Francisco" en Sinaloa. Por lo tanto, es importante tener en cuenta que la ambigüedad no es absoluta desde una perspectiva, sino también desde el punto de vista de su uso.

[14] Un diccionario consiste en la utilización de las distintas definiciones de un término.
[15] Se refiere al nivel de descomposición o grado en que pueden ser divididos los contenidos (topónimos)
[16] Disponible en: {GeoNames: http://www.geonames.org/, GeoPlanet: http://www.geoplanet.vd.ch/, WordNet: http://wordnet.princeton.edu/ }

**Tabla 3.1. Topónimos más ambiguos en GeoNames, GeoPlanet y WordNet.**

| **GeoNames** | | **GeoPlanet** | | **WordNet** | |
|---|---|---|---|---|---|
| ***Topónimo*** | ***# de lugares*** | ***Topónimo*** | ***# de lugares*** | ***Topónimo*** | ***# de lugares*** |
| San Antonio | 1,536 | Rampur | 319 | Victoria | 5 |
| Mill Creek | 1,529 | Fairview | 250 | Aberdeen | 4 |
| Spring Creek | 1,483 | Midway | 233 | Columbia | 4 |
| San José | 1,360 | San Antonio | 227 | Jackson | 4 |
| Dry Creek | 1,269 | Benito Juarez | 218 | Avon | 3 |
| Santa Rosa | 1,185 | Santa Cruz | 201 | Columbus | 3 |
| Bear Creek | 1,086 | Guadalupe | 193 | Greenville | 3 |
| Mud Lake | 1,073 | San Isidro | 192 | Bangor | 3 |
| Krajan | 1,030 | Gopalpur | 186 | Salem | 3 |
| San Francisco | 929 | San Francisco | 177 | Kingston | 3 |

En la tabla 3.2 se muestran el número de topónimos ambiguos de México de acuerdo a GeoNames y éstos son obtenidos a partir de la tabla 3.1, por que son el resultado de los topónimos más ambiguos y son topónimos mexicanos.

**Tabla 3.2. Topónimos más ambiguos de México.**

| **Topónimo** | **# de lugares** |
|---|---|
| San José | 1,135 |
| San Antonio | 867 |
| Guadalupe | 632 |
| San Francisco | 593 |
| San Isidro | 422 |
| Santa Cruz | 364 |
| Santa Rosa | 321 |
| Victoria | 200 |
| Benito Juárez | 140 |

## 3.2. Desambiguación de topónimos usando densidad conceptual

Utilizando WordNet como recurso para la GIR no se limita a su uso como "*repositorio de sentidos*" de topónimos. Los datos estructurados pueden ser explotados para adoptar los algoritmos basados en WSD al problema de desambiguación de topónimos usando WordNet. Uno de los algoritmos es el algoritmo de densidad conceptual (*Conceptual Density, CD*) introducido por Aguirre y Rigau [42], como una medida de la correlación entre el sentido de una palabra dada y su contexto. Se calculan sub-jerarquías en WordNet, determinadas por la relación de hiperonimia. El algoritmo de desambiguación por medio de CD consta de los pasos siguientes:

1. Seleccionar la siguiente palabra ambigua *w*, con |*w*| sentidos
2. Seleccionar el contexto $c_{w}$, es decir, una secuencia de palabras para *w*
3. Construir |*w*| sub-jerarquías, una por cada sentido de *w*
4. Por cada sentido *s* de *w*, calcular $CD_s$
5. Asignar a *w* el sentido que maximiza $CD_s$

En el trabajo [5], se modificó la fórmula original de densidad conceptual utilizada para calcular la densidad de sub-jerarquías *s* en WordNet con el fin de obtener un rango de frecuencias *f* [43], obteniendo como resultado la fórmula (3).

$$\mathrm{CD}\ \ \mathrm{m,f,n}\ = m^{\propto}\ \frac{m}{n}\ log^{f} \qquad (3)$$

De la fórmula (3) *m* representa el conteo de synsets[17] relevantes que contiene la sub-jerarquía, *n* representa el número total de synsets en la sub-jerarquía y *f* es el rango de frecuencia del sentido de las palabras relacionado con la sub-jerarquía (por ejemplo, 1 para el sentido más frecuente, 2 para un segundo, etc.). La inclusión del rango de frecuencias significa que los sentidos menos frecuentes son seleccionados solo cuando $m/n \geq 1$. Tanto los synsets relevantes como los synsets correspondientes al significado de una palabra eliminan la ambigüedad en el contexto de las palabras [5].

[17] Palabras agrupadas en conjuntos de sinónimos cognitivos.

### 3.2.1. Densidad conceptual y desambiguación del sentido de las palabras

La densidad conceptual trata de proporcionar una base para la medición de la cercanía de significado entre palabras, teniendo como referencia a una red jerárquica estructurada. La distancia conceptual entre dos conceptos se define en [51] como la longitud de la trayectoria más corta que conecta los conceptos en una red semántica jerárquica. En una enfoque similar [52], emplea la noción de distancia conceptual entre nodos de la red con el fin de mejorar la precisión en la indexación de documentos. En [53] se captura la similitud semántica (estrechamente relacionado con densidad conceptual) por medio de la información contenida de los conceptos en una red jerárquica. En general, estos enfoques se centran en los sustantivos (nombres). La medida de densidad conceptual entre los conceptos que se están buscando deben ser sensibles a:

- La longitud del camino más corto que conecta los conceptos involucrados.
- La profundidad en la jerarquía: los conceptos en una parte profunda de la jerarquía deben ser clasificado más de cerca.
- La densidad de los conceptos en la jerarquía: los conceptos en una parte densa de la jerarquía están relativamente más cerca que los de una región más escasa.
- La medida debe ser independiente del número de conceptos que se miden.

En [42] se han experimentado varias fórmulas para seguir los cuatro criterios expuestos anteriormente.

Para ilustrar como la densidad conceptual puede ayudar a eliminar la ambigüedad de las palabras, la figura 3.1 muestra este hecho. La palabra *W* tiene cuatro sentidos y varias palabras de contexto. Cada sentido se la palabra pertenece a una sub-jerarquía de WordNet. Los puntos en la figura 3.1, representan los sentidos de la palabra (*W*) a desambiguar o los sentidos de la palabra en el contexto. La fórmula de densidad conceptual dará mayor densidad a la sub-jerarquía que contenga más sentidos. El sentido de *W* contenido en la sub-jerarquía con la CD más alta será elegido.

Dado un concepto *c*, en la parte superior de un sub-jerarquía y dado un *nhyp* (número medio de hipónimos por nodo), la densidad conceptual para *c* cuando su sub-jerarquía contiene un número *m* (marcas) de los sentidos de las palabras para eliminar la ambigüedad está dada por la fórmula (4).

$$\mathrm{CD}\ (\mathrm{c},\mathrm{m}) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendants_i} \qquad (4)$$

La fórmula (4) muestra los parámetros calculados experimentalmente. El 0.20 trata de suavizar la exponencial *i*, *m* varía entre 1 y el número total de sentidos en WordNet. Varios valores fueron probados en [42] para obtener el parámetro y se encontró que el mejor rendimiento se logró cuando la constante del parámetro fue cerca de 0.20.

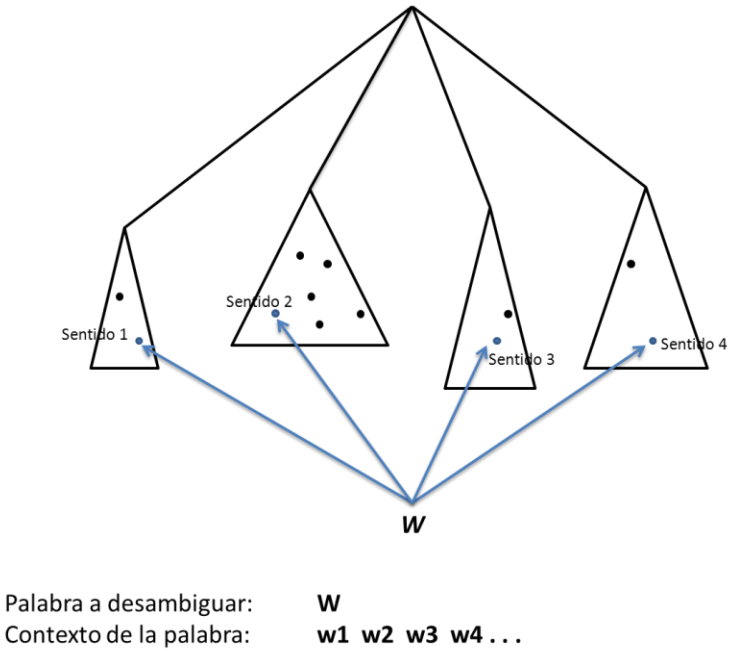


**Fig. 3.1. Sentidos de una palabra en WordNet [42].**

Para ejemplificar este proceso, se tomará como base la siguiente frase en inglés: *"The jury praised the administration and operation of Atlanta Police Department"*, se toma como ejemplo una frase en inglés, debido a que, WordNet es una base de datos léxica en inglés. Se tienen las palabras: *jury*, *administration*, *operation* y *Police Department*, son ambiguas, y la palabra a desambiguar con densidad conceptual fue "*operation*" entonces se siguen los siguientes pasos y tenemos como resultado la figura 3.2.

1) Crear una jerarquía (red) de los sustantivos en el contexto, sus sentidos y sus hiperónimos.

2) Calcular la densidad conceptual de los conceptos resultantes (sub-jerarquías).

3) El concepto más alto con la densidad conceptual se selecciona.

4) Seleccionar los sentidos por debajo del concepto seleccionado como el sentido correcto de las palabras respectivas.

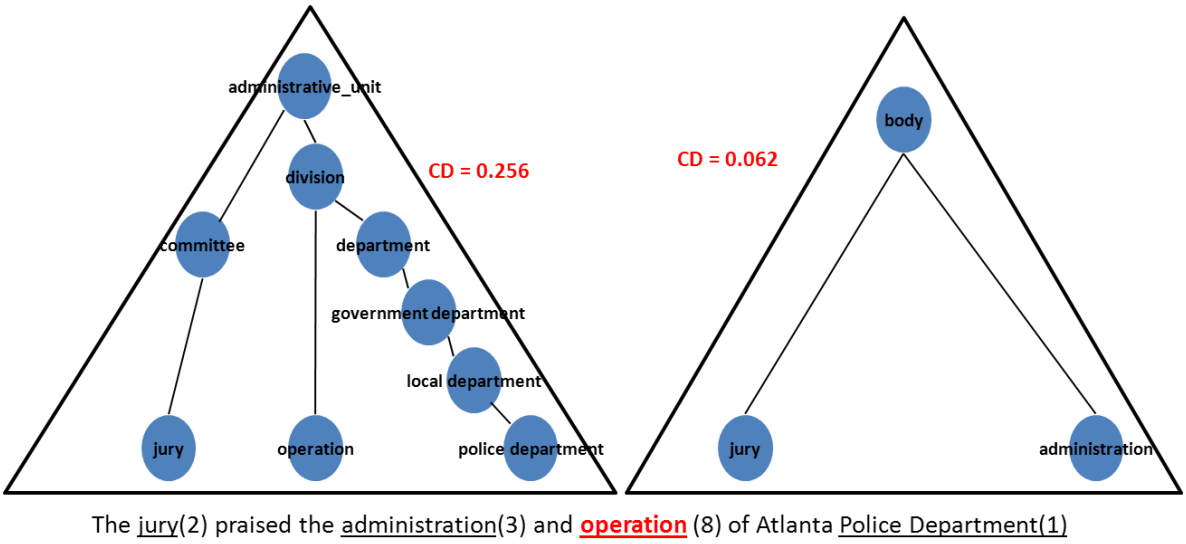


**Fig. 3.2. Ejemplo de densidad conceptual [42].**

## 3.3. Aplicaciones para la desambiguación de topónimos

La mayoría de las aplicaciones presentadas en el capítulo anterior (Estado del arte) pueden ser consideradas como aplicaciones relacionadas con el proceso de recuperación de información de una colección de textos o en otras palabras, lo que comúnmente se conoce como Recuperación de Información. Un estudio general de los módulos y las fases que constituyen el proceso de IR ha sido dado por Baeza-Yates y Ribeiro-Neto [57] y se muestra en la figura 3.3.

El paso básico en el proceso de IR consiste en tener una colección de documentos disponibles (base de datos de texto). El documento se analiza y se transforma, por medio de operaciones de texto. Una transformación típica llevada a cabo en el proceso de IR es derivada de Witten [58], que consiste en la transformación de la palabra a su forma base o forma raíz. Por ejemplo: para la palabra "geográfico", tenemos las siguientes transformaciones, "geógrafo", "geográficos", todo se reduciría a la misma raíz, "geográfico". Otra operación de texto común es la eliminación de *stopwords* (palabras vacías), con el objetivo de filtrar las palabras que generalmente no son consideradas informativas (por ejemplo, los pronombres personales, artículos, entre otras). Junto con estas operaciones básicas, el texto puede ser transformado en casi todos los aspectos que se consideren útiles para el programador de un sistema IR o método. Por ejemplo, los documentos se pueden dividir en pasos o información que no está incluida en los documentos y se pueden unir al texto (por ejemplo, si un lugar está contenido en alguna región). El resultado de las operaciones de texto constituye la vista lógica de la base de datos de texto, que se utiliza para crear el índice como resultado de un proceso de

indexación. El índice es la estructura que permite la búsqueda rápida más grande volúmenes de datos.

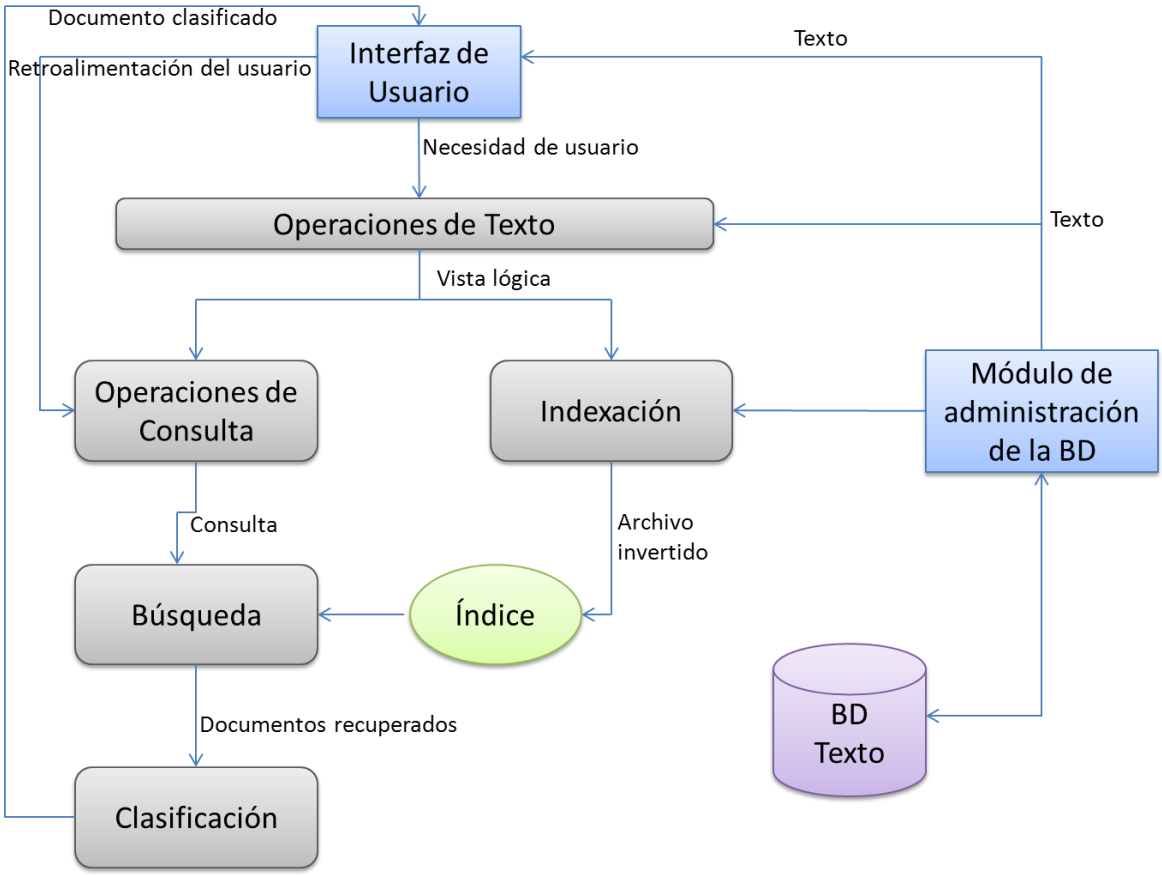


**Fig. 3.3. Visión general del proceso de recuperación de la información.**

En este punto, es posible iniciar el proceso de IR por un usuario que específica la necesidad de usuario (*user need*) que es transformada usando las mismas operaciones de texto usadas en la indexación de la base de datos textual. El resultado es una consulta que es la representación del sistema de la necesidad del usuario, aunque el término se usa a menudo para indicar lo que el usuario necesita. La consulta es procesada para obtener los documentos recuperados, que se clasifican de acuerdo a la probabilidad o relevancia (*relevance*).

Para el cálculo de relevancia, los sistemas IR primero asignan pesos a los términos contenidos en los documentos. El peso del término representa la importancia que tiene el término en el documento. Muchos esquemas de ponderación se han propuesto, pero el más conocido y probablemente más utilizado es el esquema *tf* . *idf*. El principio base de este sistema de ponderación es que un término que es frecuente en un documento dado pero

poco frecuente en la recolección debe ser particularmente informativo para el documento. Más formalmente, el peso de un término $t_i$ en un documento $d_j$ se calcula de acuerdo con el esquema de ponderación *tf.df* de acuerdo a la fórmula (5) [57].

$$w_{i,j} = f_{i,j} \text{ x } \log\frac{N}{n_i} \tag{5}$$

De la fórmula (5), $N$ es el número total de documentos en la base de datos, $n_i$ es el número de documentos en los que aparece el término $t_i$ y $f_{i,j}$ es la frecuencia normalizada del término $t_i$ en el documento $d_j$ y esta dada por la fórmula (6).

$$f_{i,j} = \frac{freq_{i,j}}{max_l \ freq_{l,j}} \tag{6}$$

De la fórmula (6), $freq_{i,j}$ es la frecuencia cruda de $t_i$ en $d_j$ (es decir, el número de veces que el término $t_i$ es mencionado en $d_j$). La parte $\log\frac{N}{n_i}$ de la fórmula de $w_{i,j}$ es la frecuencia inversa del documento $t_i$.

Los pesos se utilizan para determinar la importancia de un documento con respecto a una consulta determinada. Muchos modelos se han propuesto en este sentido, siendo el más común el modelo de espacio vectorial introducida por Salton y Lesk [59]. En este modelo, tanto la consulta como el documento se representan con un vector de *T-dimensional* (siendo $T$ el número de términos en la colección de texto indexado) que contiene sus pesos: vamos a definir $w_{i,j}$ como el peso del termino $t_i$ en el documento $d_j$ y $w_{i,q}$ como el peso del término $t_i$ en la consulta $q$, entonces $d_j$ puede ser representado como el vector $d_j = (w_{1,j}, w_{2,j}, \ldots, w_{T,j})$ y $q$ como el vector $q = (w_{1,q}, w_{2,q}, \ldots, w_{T,q})$ . En el modelo de espacio vectorial. La relevancia se calcula como una medida de similitud entre el coseno del vector del documento y el vector de consulta, se calcula mediante la fórmula (7).

$$sim(d_j, q) = \frac{d_j \,.\, q}{|d_j| \text{ x } |q|} = \frac{\sum_{i=1}^{T} w_{i,j} \text{ x } w_{i,q}}{\sqrt{\sum_{i=1}^{T} w_{i.j}} \text{ x } \sqrt{\sum_{i=1}^{T} w_{i,q}}} \tag{7}$$

Los documentos clasificados se presentan al usuario (usualmente una lista de *snippets*, que son fragmentos que están compuestos por el título y un resumen del documento) que pueden ser utilizados para retroalimentar (*feedback*) la mejora de los resultados en el caso de no estar satisfechos con estos resultados. La evaluación de los sistemas de IR se lleva a cabo mediante la comparación de una lista de resultados con una lista de documentos relevantes y no relevantes recopilados por los evaluadores humanos.

## 3.4. Desambiguación de topónimos en recuperación de información geográfica

La ambigüedad léxica y su relación con la IR ha sido objeto de muchos estudios en la década pasada. Uno de los temas más debatidos ha sido la desambiguación del sentido de las palabras (WSD) [44], puede ser útil para IR o no. Mark Sanderson ha investigado a fondo el impacto de la WSD en IR.

En [45, 46] Sanderson experimentó con pseudo-palabras (palabras ambiguas creadas artificialmente), demostrando que cuando se introduce la ambigüedad se desambigua con una precisión de 75% (25% error). El argumentó que solo la precisión alta (superior al 90%) en WSD puede permitir obtener un rendimiento beneficioso y mostró también que el uso de la desambiguación es útil solo en el caso de consultas cortas, debido a la falta de contexto. Más tarde, Gonzalo [47], llevó a cabo algunos experimentos de IR con el corpus SemCor, encontrando que las tasas de error por debajo del 30% producen mejores resultados que la palabra estándar de indexación. Más recientemente, Stokoe [47] fue capaza de obtener una mayor precisión en IR, utilizando un desambiguador de WSD con una precisión de 62.1%; en sus conclusiones, el afirma que los beneficios de usar WSD en IR se pueden presentar dentro de ciertos tipos de recuperación o en escenarios de recuperación específicos.

GIR constituye un escenario de recuperación, dado que la asignación de un referente topónimo válido puede alterar significativamente los resultados de una consulta determinada. Por ejemplo, la devolución de resultados en una búsqueda: Salamanca puede regresar resultados de Salamanca, España cuando en realidad se esperaban resultados de Salamanca, Guanajuato.

Algunos trabajos de investigación sobre procesamiento de lenguaje natural (*Language Natural Processing, PLN*) tienen varios errores sobre el rendimiento de GIR, esto se ha llevado a cabo en el trabajo de Stokes [48]. Su sistema experimental utiliza los motores de búsqueda del Zettair[18] con un índice ampliado, añadiendo jerarquías basadas en geo-términos dentro del índice como si fueran "palabras", una técnica para la cual no es necesario introducir estructuras de datos espaciales. Por ejemplo, "Melbourne, Victoria" en el índice con el término "@OC-Australia-Victoria-Melbourne" (OC significa "Oceanía").

Sus experimentos mostraron que el bajo recuerdo (*recall*) tiene un mayor impacto en la eficacia de la baja precisión del Reconocedor de Entidades (*Named Entity Recognizer*, *NER*) y que estadísticamente disminuye el puntaje en el MAP (*Mean Average Precision*) cuando la precisión de desambiguación se reduce de 80% a 40%. Sin embargo, el carácter personalizado y el pequeño tamaño de la colección no permiten generalizar los resultados.

[18] Disponible en: http://www.seg.rmit.edu.au/zettair/

### 3.4.1. Recuperación de información geográfica

La Recuperación de Información Geográfica se encarga de proporcionar accesos a fuentes de información geográfica y espacial. Esta incluye todas las áreas que tradicionalmente han sido el núcleo de la investigación en IR, con un énfasis o extensión hacia aspectos geográficos y espaciales, tanto en procesos de ponderación, almacenamiento y de indexado [54]. GIR es un tema interdisciplinario de reciente nacimiento, pero de rápido desarrollo académico y comercial. En donde se procesan datos por medio de alguna noción de la relevancia geográfica en la información. La investigación actual enfrenta varios retos, de entre los cuales se destacan:

- La extracción de términos geográficos de datos estructurados, y de reto aun mayor, de datos no estructurados.

- La identificación y eliminación de ambigüedades en los procedimientos de extracción.

- Metodologías para almacenamiento eficiente de ubicaciones y sus relaciones.

- Desarrollo de máquinas de búsqueda y algoritmos para aprovechar las características de la información geográfica.

- La combinación de relevancia geográfica y contextual para aportar una relevancia significativa a los documentos; así como, técnicas para permitir al usuario interactuar y explorar los resultados de consultas para sistemas GIR. En este punto, observar figura 3.4 para distinguir lo que IR y recuperación de datos.

En IR el modelo subyacente que proporciona acceso a los documentos[19] es probabilista. Se interesa con cuestiones subjetivas e indeterminadas. Es decir, si un documento es relevante (a un cierto grado) para el usuario y su consulta. Mientras que la recuperación de datos es determinista con respecto a operaciones de recuperación. Si un documento cumple las condiciones especificadas en una consulta, entonces éste es por definición "relevante".

En GIR se enfoca en ambos aspectos: recuperación determinista (por ejemplo, encontrar todos los conjuntos de datos que contienen información sobre una coordenada particular) y recuperación probabilística (tales como encontrar todos los municipios cerca de un río) [55, 56]. GIR se enfoca en los problemas de encontrar fuentes de información que están relacionadas con ubicaciones geográficas. Debido a que la mayoría de maquinas de búsqueda tratan la terminología geográfica al igual que otra terminología, lo cual trae como consecuencia que se recuperen documentos irrelevantes. Actualmente, GIR trabaja en los siguientes retos:

- Existen muchos lugares con el mismo nombre (documentos que se refieren al lugar equivocado son recuperados).

- Existen muchos usos para nombres de lugares. Por ejemplo, se usan nombres de personas y de organizaciones para referirse a un lugar.

- Existen consultas que incluyen preposiciones espaciales, tales como cerca, afuera, que requieren consideraciones especiales en el ámbito geográfico. Por ejemplo técnicas de geo-parsing[20] y geo-coding[21].

Una de las metas en GIR es la de proporcionar acceso a fuentes de información georeferenciada. Por lo tanto, podemos considerar a la recuperación de información geográfica como una especialización de la recuperación de información clásica. Esto incluye todas las áreas que tradicionalmente conforman el núcleo de la investigación en IR, con especial énfasis en el indexado y recuperación de información orientada espacial y geográficamente. Estos procesos se realizan de manera común a través de consultas espaciales (*Spatial Query*) o de consultas geográficas.

En la literatura computacional han existido distinciones entre IR y recuperación de datos (*data retrieval*) este último término está asociado con sistemas manejadores de bases de datos (SMDB). En la práctica esta distinción es más de grado que de tipo. En la figura 3.4 se muestra un espectro de varios atributos relacionados con la recuperación de información y la recuperación de datos. El estudio de estos atributos permitirá establecer la diferencia entre estos.

Como se observa en la figura 3.4 el modelo utilizado en IR para el acceso a los documentos es *probabilista*. Está referido con cuestiones subjetivas e indeterminadas como pueden ser: que un documento satisfaga, *en cierto grado*, las necesidades de información de un usuario (que sea relevante para el usuario de acuerdo a su consulta-requerimiento). Por otro lado, la recuperación de datos, es *determinista* considerando operaciones de recuperación. Es decir, si un documento cumple las condiciones especificadas en la *consulta* del usuario, entonces es por definición "relevante".

Entonces, cuando nos referimos a recuperación de información geográfica hablamos de ambos tipos de modelos de recuperación tanto determinista (por ejemplo: encontrar todos los conjuntos de datos que contienen "cualquier" información para una coordenada en particular) como probabilista (por ejemplo: encontrar todas las ciudades cerca de un gran río). De esta manera, para recuperar información se requiere de un indexado tanto para garantizar un acceso eficiente a grandes bases de datos, como para organizar y limitar el conjunto de elementos de la base de datos que sean accesibles.

---

[19] Se utiliza el término documento para representar cualquier elemento de interés potencial en una colección o base de datos, sin considerar el contenido -- texto, imágenes, mapas, video, etc. -- o la forma – papel o digital.
[20] El propósito consiste en detectar terminología geográfica.
[21] Consiste en adjuntar una referencia única de ubicación para un documento.


**Fig. 3.4. Espectro: Recuperación de información vs. Recuperación de datos [54].**

La mayoría de los sistemas IR obtienen sus elementos índices a partir del contenido de los elementos a ser indexados. Esta obtención puede ser una simple extracción (tal como extraer palabras clave, "keywords", de un texto) o extracción por inferencia (por ejemplo: el mapeo de una palabra a los términos en un tesauro) o podría ser un análisis inteligente de asignación de elementos índice (por ejemplo: asignar encabezados, "títulos", de los temas en un documento). Mientras que en la recuperación de datos, el elemento por sí mismo, en su totalidad, es la unidad de indexación. Claro, que en la escala de la figura 3.4 esto no es lineal o continuo dado que ambos tipos de indexado podrían estar presentes dentro del mismo sistema. En GIR ambos extremos de la escala están mezclados. Por ejemplo, usando indexado inteligente (por ejemplo: asignar las coordenadas de un cuadro de selección a una fotografía aérea), e indexado por inferencia (asignar coordenadas a los lugares que son mencionados en un texto).

En la recuperación actual de elementos de una base de datos, los algoritmos utilizados para *coincidir*[22] una consulta con los elementos del índice (o el contenido de una base de datos)

[22] El término coincidencia será utilizado como el equivalente al término matching el cual es ampliamente usado en las áreas de lenguajes de programación, bases de datos, ingeniería de software, entre otras.

están basados en el modelo de recuperación tradicional. Los modelos de recuperación de información guían a una clase de algoritmos de recuperación que son probabilistas por naturaleza, y pueden involucrar el cálculo de probabilidades y el uso de métodos de inferencia estadísticos, aunque también pueden tener un enfoque basado en otro modelo del espacio de documentos.

Estos modelos están enfocados a encontrar todas las coincidencias (parciales) potenciales entre una *consulta*[23] y un documento, mientras que la ponderación de éstas trabaja con base en parámetros que miden el "grado de coincidencia", de forma tal que las "mejores" coincidencias reciben las más altas ponderaciones.

Los algoritmos de recuperación de datos son deterministas, y entonces demandan una *coincidencia* exacta entre la especificación de la *consulta* y el contenido de la base de datos. La lógica de Boole es utilizada en el procesamiento de los lenguajes basados en consultas (prácticamente en todos los sistemas comerciales manejadores de bases de datos) como también en los catálogos en línea y sistemas de recuperación de información, este también es un algoritmo determinista. En GIR la coincidencia determinista aproximada, parcial y precisa son de utilidad en el procesamiento de *consultas* espaciales y geográficas [55].

Las consultas en sistemas de IR son generalmente expresadas como enunciados en lenguaje natural de acuerdo a las necesidades de los usuarios que buscan información. Estas *consultas* son intrínsecamente imprecisas y pueden ser ambiguas. Mientras que en la recuperación de datos la *consulta* es típicamente expresada en algún tipo de lenguaje estructurado cuya sintaxis es precisa y de características semánticas.

Por lo tanto, cuando el objetivo consiste en recuperar todas las unidades de la base de datos que coinciden exactamente con las especificaciones de la *consulta*, el problema de la ambigüedad en el enunciado de la *consulta* no se presenta (respecto a lo que se busca).

En consecuencia, los tipos de consulta pueden reflejar los modelos subyacentes de los sistemas de recuperación. En la recuperación de la información las consultas son consideradas como una "pista o clave" acerca de lo que el usuario podría considerar un elemento relevante de la base de datos, y adicionalmente esta recuperación está basada en que también un elemento coincide con la "clave".

Típicamente, los resultados de una búsqueda son presentados en un orden ponderado, donde el criterio de ponderación se basa en el grado de "coincidencia" entre la *consulta* y el elemento de la base de datos.

[23] El término consulta es usado como el equivalente al término query utilizado en base de datos, aunque en este caso existen diferentes estructuras para definirlo.

En la *recuperación de datos* la consulta es tratada como una precisa especificación de los elementos deseados de la base de datos y la recuperación está basada en una correspondencia exacta entre el elemento y la *consulta*, a menos de que sea explícitamente indicado por el sistema o por el usuario como parte de la *consulta*, no hay ponderación u orden impuesto sobre los resultados para una *consulta* de recuperación de datos. Adicionalmente, la recuperación de información geográfica, es conocida como un área de investigación aplicada, la cual combina aspectos de investigación de otras áreas tales como DBMS, interfaces de usuario, GIS, y por supuesto IR. También trabaja con aspectos tales como el indexado, búsqueda, recuperación y navegación, exploración en diversas y múltiples fuentes de información *geo-referenciada*, y al diseño de sistemas que cumplan estas tareas de forma efectiva y eficiente [55].

### 3.4.2. La tarea de la recuperación de información geográfica

Se puede definir la tarea de la recuperación de información geográfica como la recuperación de documentos relevantes en respuesta a una consulta con el formato <*tema*, *localización*>, donde la relación espacial puede implicar implícitamente contenido, o explícitamente ser seleccionado de un conjunto de posibles opciones topológicas, direccionales o de proximidad [49].

Existen una amplia variedad de enfoques para resolver la tarea GIR, que van desde aproximaciones simples de recuperación de información sin indexación de términos geográficos a arquitecturas que hacen uso de técnicas de procesamiento del lenguaje natural para extraer localizaciones e información topológica de los documentos y las consultas. Algunas de las técnicas usadas en la actualidad incluyen extracción de entidades geográficas, análisis semántico, bases de conocimiento geográfico (como ontologías, tesauros o gazetteers), técnicas de expansión de consultas y desambiguación geográfica.

En la figura 3.5 se puede observar la arquitectura básica empleada en el sistema GIR GeoUJA [50]. Este sistema ha sido desarrollado SINAI[24] para resolver la tarea de la recuperación de información geográfica.

### 3.4.3. Recuperación de información geográfica por interpretación de datos

Actualmente, en GIR se ha enfocado únicamente en la recuperación de datos y no en los procesos aplicados a los mismos (interpretación de datos). Por otra parte, no es lo mismo recuperar numerosas capas de datos, que recuperar una capa de datos de forma aislada. Es decir, si se consideran los procesos que afectan a los datos en conjunto (como en el análisis espacial) se pueden obtener resultados de mayor relevancia ya que se están considerando más aspectos en la recuperación.

[24] Disponible en: http://sinai.ujaen.es

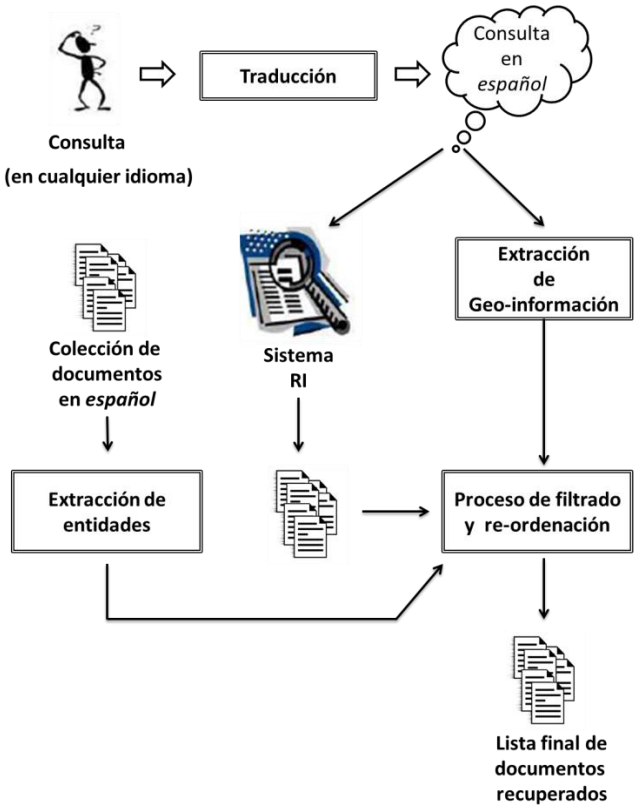


**Fig. 3.5. Arquitectura básica del sistema GIR GeoUJA [50].**

Entonces, lo que se desea resaltar es que los enfoques se han dirigido a recuperar datos para posteriormente aplicarles análisis y procesos. Pero, resultaría más útil recuperar datos basándose en dichos análisis y procesos. Por lo tanto, lograr una recuperación basado en la interpretación de los datos representa una mayor utilidad y extendería el enfoque a usuarios comunes.

Por ejemplo, si se desea conocer donde existen terrenos fértiles para el cultivo de maíz, una recuperación de datos, arrojaría capas de datos de terrenos, clima, áreas de cultivo, entre otras. En este caso el usuario tiene que tener un conocimiento en GIS para integrar e interpretar los datos.

Pero, si la recuperación se basa en interpretación, entonces se arrojan datos que ya incluyen las capas resultantes de un análisis para encontrar terrenos fértiles. En otras palabras, ***el usuario recibe la información integrada y procesada***, facilitando muchas de sus tareas. Además, si consideramos los enfoques conducidos por ontologías, en donde los datos se procesan de acuerdo a sus relaciones semánticas y a sus propiedades. Entonces, se pueden obtener resultados que con los enfoques sintácticos son omitidos. Por lo tanto, los trabajos que se dirigen en esta dirección permitirán mejorar los procesos de recuperación y consulta de datos.

### 3.4.4. Principales técnicas de PLN aplicadas en un sistema GIR

En el estudio de las principales técnicas PLN aplicadas en una arquitectura GIR, en general, todas las arquitecturas presentadas realizan un preprocesamiento tanto a las colecciones de documentos como a las consultas formuladas. Este análisis lingüístico consiste en aplicar un extractor de raíces (*stemmer*), una lista de palabras sin contenido semántico (*stopwords*), para eliminar las palabras vacías, y un Reconocedor de Entidades (*Named Entity Recognizer, NER*) para detectar y reconocer posibles entidades en cualquier texto.

Según el estudio realizado, el *stemmer* más utilizado es el Porter Stemmer[25]. También se usa en varios sistemas, pero con menos frecuencia que el anterior, el Snowball Tartarus[26]. Con respecto a la lista de stop-words para el inglés, la más utilizada ha sido la creada por Salton y Buckley[27], que consta de 571 palabras.

En relación a los reconocedores de entidades más empleados, hay sistemas que han optado por implementar sus propios reconocedores haciendo uso de distintas bases de conocimiento geográficas y tesauros [60,61], pero la mayoría han empleado Lingpipe[28] como herramienta NER.

Con respecto al análisis de los distintos sistemas, es poco habitual utilizar herramientas de etiquetado POS (Part Of Speech), aunque algunos sistemas como [60] hacen uso de un etiquetador POS estadístico llamado TnT.

Por último, otra herramienta importante en el ámbito del PLN son los traductores o sistemas de traducción automática (*Machine Translation, MT*). Para la tarea GIR es necesario utilizarlos cuando la consulta planteada y la colección a indexar están en idiomas distintos (tarea multilingüe). En [61] se hace uso del traductor *LEC Power Translator*. En nuestro sistema GIR GeoUJA [50] se utiliza un sistema propio de traducción automática llamado SINTRAM (*SINai TRAnslation Module*) [62].

En general, la arquitectura de cualquier sistema GIR parte de un modelo básico de recuperación de información. Por tanto, un elemento esencial en todos los sistemas presentados es la herramienta utilizada como motor de búsqueda.

---

Consultar:
[25] http://tartarus.org/martin/PorterStemmer/
[26] ftp://ftp.cs.cornell.edu/pub/smart/english.stop
[27] http://alias-i.com/lingpipe/
[28] http://gate.ac.uk/

## 3.5. Desambiguación de topónimos en búsqueda de respuestas

La búsqueda de respuestas, consiste en que dada una cierta cantidad de documentos no estructurados, el sistema debe ser capaz de recuperar respuestas a preguntas planteadas en lenguaje natural. QA es a veces visto como una forma particular de Recuperación de Información en el que la cantidad de información recuperada es la cantidad mínima de información que se requiere para satisfacer las necesidades del usuario. De esta definición se desprende que los sistemas de QA tienen que lidiar con problemas más complicados que los sistemas IR: en primer lugar, ¿cuál es la "mínima" cantidad de información con respecto a una pregunta determinada?, ¿Cómo debe ser extraída esta información?, ¿Cómo debe ser presentada la información al usuario? Estos son sólo algunos de los muchos problemas que se pueden encontrar. Los mejores resultados obtenidos por los sistemas QA son típicamente de entre 40 y 70 por ciento en la precisión, dependiendo del lenguaje y el tipo de ejercicio. Por lo tanto, algunos esfuerzos se han llevado a cabo con el fin de centrarse sólo en determinados tipos de preguntas (dominio restringido QA), incluyendo leyes, genómica, dominio geográfico, entre otros.

Un sistema QA generalmente se divide en tres módulos principales: Clasificación y Análisis de la Pregunta, Documento o Recuperación del Pasaje y Extracción de Respuestas. Estos módulos tienen que lidiar con diferentes desafíos técnicos que son específicos para cada fase. La arquitectura genérica de un sistema de búsqueda de respuestas se muestra en la figura 3.6.

La clasificación de preguntas (*Question Classification, QC*) se define como la tarea de asignar una clase para cada pregunta formulada a un sistema. Sus principales objetivos son permitir la extracción al módulo de respuestas a diferentes extracciones de respuesta (*Answer Extraction, AE*) para cada tipo de estrategia y para restringir las respuestas candidatas. Por ejemplo, la extracción de la respuesta a "¿Qué es vicodina?", se está buscando una definición, que no es lo mismo que la extracción de la respuesta a ¿Quién inventó la radio?, que se pide el nombre de una persona. La clase que se puede asignar a una pregunta afecta en gran medida todos los pasos siguientes de la QA y por lo tanto, es de vital importancia asignar adecuadamente.

Un estudio realizado por Moldavo [63] revela que más del 36% de los errores en QA se deben directamente a la fase de la pregunta de clasificación. Los enfoques a la pregunta de clasificación (QC) se pueden dividir en dos categorías: los clasificadores basados en modelo y clasificadores supervisados. En ambos casos, una pregunta importante es representada por la taxonomía de las clases de la pregunta que puede ser clasificada. El diseño de los sistemas QC siempre comienzan por determinar cuál es el número de clases y la manera de cómo organizarlas.

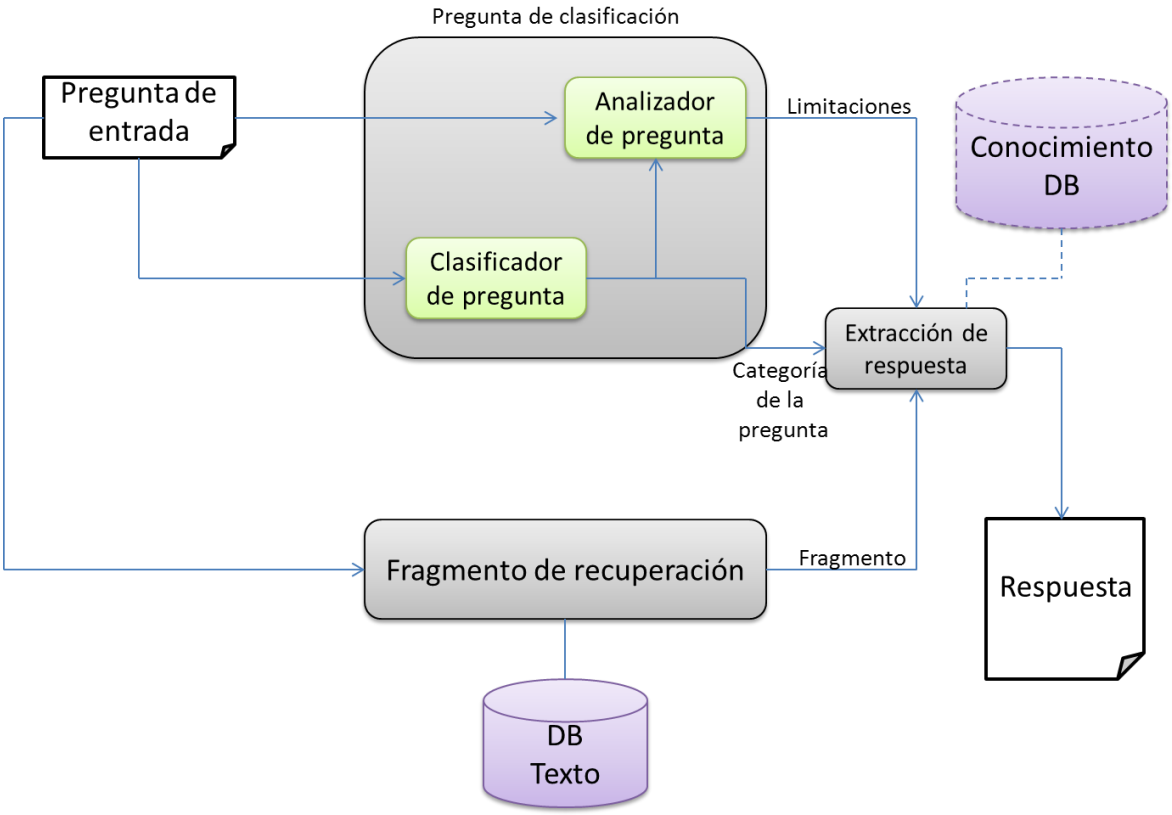


**Fig. 3.6. Arquitectura genérica de un sistema de búsqueda de respuestas.**

Otra tarea importante realizada en la primera fase es la extracción del *enfoque* y el *objetivo* de la pregunta. El enfoque es la propiedad o entidad que es buscada por la pregunta. El objetivo está representado por el acontecimiento u objeto en el cual se trata la pregunta. Por ejemplo, en la pregunta "¿Cuántos habitantes hay en Cuetzalan?", el objetivo es "habitantes" y el objetivo es "Cuetzalan". Los sistemas suelen extraer información utilizando herramientas de procesamiento de lenguaje natural, tal como etiquetadores POS (*Part-Of-Speech tagging*) y analizadores superficiales (*chunkers*).

Una recuperación de fragmentos (*Passage Retrieval, PR*) es un sistema de aplicación de IR que devuelve piezas de textos (fragmentos) que son relevantes para la consulta del usuario en lugar de devolver una lista por orden de documentos. QA-Orientados a los sistemas PR presentan algunos problemas técnicos que requieren la mejora de los métodos estándar existentes en IR o la definición de nuevos. En primer lugar, la respuesta a una pregunta puede estar relacionada con los términos utilizados en la propia pregunta, haciendo búsquedas clásicas basadas en términos buscando métodos inútiles. Estos métodos suelen buscar los documentos que se caracterizan por una alta frecuencia de términos de consulta.

Por ejemplo, en la pregunta, "¿Qué es un BMW?", el único término que no es *stopword* es "BMW" y un documento que contiene el término "BMW" muchas veces es probable que no contenga la definición de la compañía. Otro problema es determinar el tamaño óptimo del fragmento: si es demasiado pequeño, la respuesta puede no estar en el contenido del fragmento; si es demasiado largo, puede traer información que no esté relacionada con la respuesta, que requiere de un módulo de extracción de respuestas más preciso.

La fase de extracción de respuestas (*Answer Extraction*) es responsable de extraer la respuesta de los fragmentos. Cada pieza de información extraída durante las fases anteriores es importante a fin de determinar la respuesta correcta. El principal problema que se puede encontrar en la fase es el de determinar ¿cuál de las posibles respuestas es correcta o la más informativa?

Por ejemplo, la respuesta para "¿Qué es BMW?" puede ser "Un fabricante de automóviles", sin embargo, las mejores respuestas podrían ser "Un fabricante de automóviles alemán" o "Un fabricante de autos de lujo y deportivos con sede en Múnich, Alemania"

Otro problema que es similar al anterior, está relacionado con la normalización de cantidades: la respuesta a la pregunta "¿Cuál es la distancia de la Tierra al Sol?" puede ser "149, 597,871 km", "una unidad astronómica (AU)", "92, 955,807 millas" o "casi 150 millones de kilómetros". Estas son las descripciones de la misma distancia y el módulo de extracción de respuestas debe tener esto en cuenta para explotar la redundancia.

# CAPÍTULO 4.
# UNA METODOLOGÍA PARA LA DESAMBIGUACIÓN DE TOPÓNIMOS

En este capítulo se establece el marco metodológico de la presente investigación, incluye una descripción de cada uno componentes de la metodología propuesta y como estos fueron empleados para la desambiguación de topónimos.

# Capítulo 4

# Una metodología para la desambiguación de topónimos

En este capítulo se describe la metodología que se propone para resolver la tarea de desambiguación de topónimos. En primer término, se da una visión general de la metodología y de sus componentes. La arquitectura propuesta se estructura en tres módulos que agrupan respectivamente a componentes que intervienen en el flujo de trabajo para la desambiguación de topónimos. La metodología se centra principalmente en la utilización de una ontología como repositorio de sentidos, un corpus enriquecido como contexto y un método desambiguador basado en un modelo de clasificación.

## 4.1. Descripción de la metodología propuesta

Una arquitectura completa para la tarea de desambiguación de topónimos debe estar centrada en torno a un repositorio de sentidos; esto quiere decir la utilización de un diccionario, tesauro u ontología. En esta metodología se utilizará una ontología, donde se indiquen los distintos sentidos de los topónimos. Además, debe proporcionar un contexto y un método desambiguador, capaz de identificar los topónimos ambiguos y poder desambiguarlos. Finalmente, se evalúa la propuesta observando resultados favorables. La figura 4.1 ilustra la metodología propuesta para la tarea de desambiguación de topónimos. La arquitectura se divide en tres capas: el repositorio de sentidos, el contexto y el método desambiguador. La parte que corresponde al repositorio de sentidos (Sección 4.2) de la figura 4.1 muestra los componentes que intervienen en la utilización de la ontología construida. En la Sección 4.3 se describe el contexto, en la figura 4.1 se observan los cuatro componentes que forman esta capa: el etiquetado del corpus con el etiquetador Tree Tagger[29], se extraen los términos multipalabra mediante *n_gramas*, se detectan esos términos en el corpus con la ayuda de la ontología y se hace un etiquetamiento manual para de esta forma tener un corpus etiquetado manualmente que ayudará para la siguiente capa que es el desarrollar el método desambiguador (Sección 4.4).

[29] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger

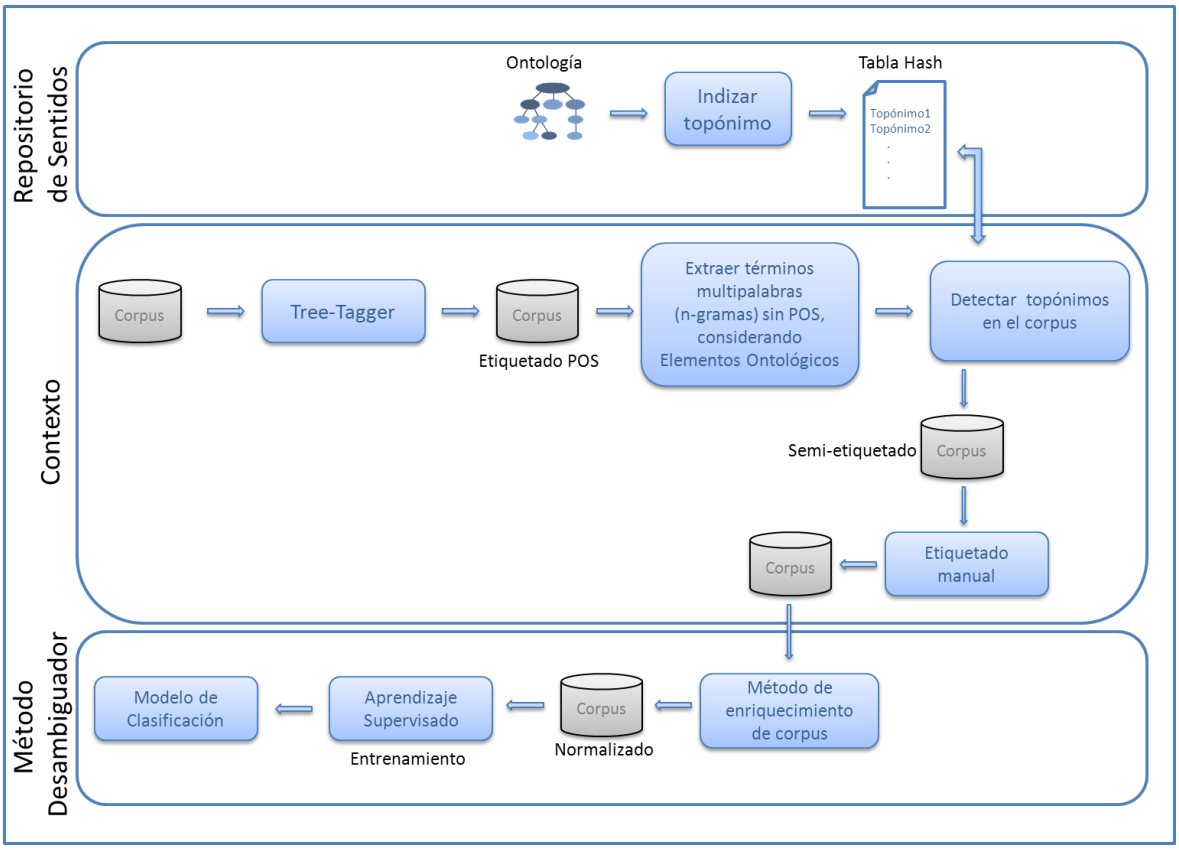


**Fig. 4.1. Metodología general para la desambiguación de topónimos.**

El repositorio de sentidos servirá para obtener los referentes candidatos a topónimos, una vez obtenido el nombre del lugar hay que desambiguarlo, ya que, muchos de los topónimos existentes son compartidos por varios lugares. Es decir, los topónimos pueden ser ambiguos y pueden tener algunos de los dos tipos de ambigüedad GEO/GEO o GEO/NO-GEO.

La ambigüedad de topónimos tipo GEO/GEO es la que se produce cuando dos localizaciones tienen el mismo nombre. En el caso de la ambigüedad tipo GEO/NO-GEO se da cuando existe confusión entre un topónimo y un término que no lo es. Por ejemplo, cuando en un texto "Benito Juárez" hace referencia al presidente "Benito Juárez" y no a la ciudad. Para ambos casos, los dominios de aplicación son la extracción y recuperación de información. En esta metodología, se aborda la ambigüedad tipo GEO/NO-GEO, debido a que considera características espaciales asociada estrechamente a la naturaleza espacial de la información con la que se trabaja.

La metodología propuesta para la desambiguación de topónimos produce resultados útiles al utilizar colecciones de noticias en Español como corpus, al hacer el enriquecimiento de este con información descargada de la Web se logra incrementar la precisión del método.

## 4.2. Construcción de una ontología como recurso de apoyo en la desambiguación de topónimos

En esta sección se presenta una ontología espacial que sirve como repositorio de sentidos para una de las tareas importantes de la recuperación de información geográfica como lo es la desambiguación de topónimos para el idioma Español en particular para la República Mexicana. La ontología incluye la representación de objetos geográficos naturales y artificiales. El uso de ontologías ayuda a solventar problemas tales como encontrar el verdadero sentido de las palabras, incluyendo en un solo repositorio todos los conceptos y relaciones del dominio de trabajo. De esta manera se evitan errores en el manejo de la información y es posible unificar el lenguaje de la comunicación en función de sus diferentes sentidos semánticos para desambiguar. La ontología espacial presentada fue desarrollada en el gestor de ontologías Protégé y posteriormente validada con el razonador RacerPro para garantizar la consistencia de la misma.

### 4.2.1. Ontologías

Las ontologías permiten que las máquinas puedan intercambiar información de forma efectiva y eficiente. Para ello proporcionan formalismos y estructuran la información permitiendo un cierto grado de razonamiento automático. Gruber [64] creó una de las definiciones más citadas del concepto de ontología en el ámbito de la informática: "una especificación explícita y formal sobre una conceptualización compartida".

Gruber [65] es también uno de los autores más citados al identificar los cinco componentes básicos del modelado de ontologías, los cuales se enuncian a continuación:

- Conceptos. Son las ideas básicas que se intentan formalizar.
- Relaciones. Representan las interacciones y los enlaces entre los conceptos del dominio.
- Funciones. Son casos especiales de relaciones donde se identifican elementos mediante el cálculo de una función que considera varios elementos de la ontología.
- Instancias. Se usan para representar elementos determinados en una ontología.
- Axiomas. Los axiomas formales sirven para modelar sentencias que son siempre ciertas. Normalmente, se usan para representar conocimiento que no puede ser formalmente definido por los componentes descritos anteriormente consiguiendo así una mayor capacidad expresiva del dominio. Además, también se usan para verificar la consistencia de la propia ontología.

### 4.2.2. Ontologías Espaciales

Las Ontologías Espaciales, de acuerdo a [66], son una extensión de la Lógica Descriptiva (DL), con un dominio concreto para la dimensión espacial (es decir, considera objetos espaciales tales como puntos, líneas, polígonos), para así permitir la combinación de representación del conocimiento y el razonamiento espacial dentro de un paradigma único. El dominio concreto está definido por un conjunto de predicados representando relaciones topológicas entre objetos. La habilidad de definir roles topológicos facilita la especificación de conceptos y objetos espaciales. También provee acceso algoritmos de razonamiento espacial que permiten la extensión del razonamiento terminológico de la dimensión espacial.

Algunos investigadores, como Spaccapietra [67], dividen esta Ontología en espacio y tiempo; las ontologías de tiempo definen los conceptos que son usados en un tiempo especificado y elementos temporales, como son las instancias, intervalos, cronómetros, entre otros, y relaciones temporales como son precedentes, antecedentes, entre los más significativos, pero de igual manera son denominados espaciales. Debido a la aparición de una gran cantidad de información geográfica y mapas en Internet, de casi todos los sitios posibles sobre la tierra, aparecen los Servicios de la Web Semántica [68] que convienen a un tipo de tecnologías más elaboradas, en un mundo donde se cree que más del 80% de los datos tiene un componente geográfico, como lo son las nuevas aplicaciones de mapas publicadas en el Web. Los mapas web muestran recientemente grandes crecimientos, su integración dentro del dominio espacial aparece como un paso esencial hacia la adopción de la tecnología SWS (*Shore Wireless Service*). Sin embargo, el espacio geográfico como un único pero total dominio encuadrado tiene especificaciones que describen semánticas más reconocidas. Además, los Sistemas de Información Geográfica necesitan adoptar habilidades humanas cognitivas de representación espacial y razonamiento.

La aparición de las Ontologías Espaciales sirve como soporte a este tipo de tecnologías para acceder y compartir información utilizando como componente esencial a los aspectos espaciales y temporales.

### 4.2.3. Ontología: Desarrollo y Descripción

El desarrollo de la ontología geográfica propuesta en este trabajo de tesis, se enmarca en un proyecto global ambicioso, cuyo objetivo es de ser tomado como repositorio de sentidos para resolver la tarea de desambiguación topónimos. Cabe mencionar que esta ontología puede ser utilizada para desambiguar topónimos en consultas a la Web, relacionadas por ejemplo con: el ejercicio y la promoción del turismo en México, aplicaciones de cambio climático, realización de planeaciones urbanas y desarrollo de planes estratégicos de tipo económico-sociales entre otras. Dentro de este proyecto, uno de los aspectos a desarrollar es la implementación de una ontología que permita identificar el topónimo correcto en un contexto (corpus) y a su vez mediante la ontología obtener su posición geográfica (latitud y longitud).

Para cumplir con esta meta, se presenta una ontología espacial que describe el espacio geográfico considerando los objetos geográficos naturales y artificiales, entendiendo por objetos naturales aquellos que fueron creados por la naturaleza (bahía, golfo, lago, etc.) y por objetos artificiales los que son creados por el hombre (puente, aeropuerto, ferrocarril, etc.). Hasta este momento la ontología propuesta sólo comprende objetos geográficos de la República Mexicana, debido a que en otros países de habla hispana la división política varía de acuerdo a cada país y no se garantizaba la consistencia de la ontología durante el proceso de validación con en el razonador lógico RacerPro.

Existen varios lenguajes ontológicos para implementar ontologías, los cuales proporcionan distintos niveles de formalismo y facilidad de razonamiento. El lenguaje OWL[30] (*Ontology Web Language*), estandarizado por el W3C (*World Wide Web Consortium*), permite definir ontologías con varios niveles de detalle. Dicho lenguaje se puede categorizar en tres especies o sublenguajes: OWL-Lite, OWL-DL y OWL-Full.

La ontología espacial se implementó en Protégé[31] empleando el sublenguaje OWL-DL, debido a que está diseñado para aquellos usuarios que requieren máxima expresividad conservando completitud computacional (se garantiza que todas las conclusiones sean computables) y resolubilidad (todos los cálculos se resolverán en un tiempo finito). Una de las ventajas de utilizar Protégé es que cuenta con el manejo de instancias sobre las clases, así como restricciones para generar éstas. En la tabla 4.1, se muestran las características generales de la ontología espacial implementada en el desarrollo de la metodología para la desambiguación de topónimos.

#### 4.2.3.1. Desarrollo de la Ontología

En el desarrollo de ontologías, el primer paso es identificar la información que se quiere representar. Lo más adecuado es tomar como base de conocimiento de expertos en el dominio en cuestión, aprovechando posibles categorizaciones o clasificaciones ya existentes.

Para la definición de aspectos genéricos de la ontología, han servido como base la ontología presentada en [40], una ontología mixta que proporciona un vocabulario de clases y relaciones para describir un área específica. En este caso el espacio geográfico incluye un análisis de la distribución de 500 millones de hispanohablantes estimados en el mundo.

**Conceptos (Jerarquía de Clases)**
La ontología se desarrolla a partir de la jerarquía de clases que se muestran en la figura 4.2. En esta figura se pueden distinguir tres clases de alto nivel: *Extension_Geografica*, *Estructura_Geografica* y *Localizacion*.

[30] www.w3.org/TR/owl-features
[31] protege.stanford.edu

**Tabla 4.1. Características generales de la Ontología Espacial.**

| Característica | Nombre | Descripción |
|---|---|---|
| Jerarquía de clases | ExtensionGeografica<br>EstructuraGeografica<br>Localizacion | Las clases constituyen las unidades básicas de la ontología que se pretende formalizar |
| Propiedades de objeto | contieneLocalizacion<br>formaParteDeLocalizacion<br>tienePuntoDeInicio<br>tienePuntoFinal<br>estaADistancia | Relacionan todas las clases de la ontología con la entidad Localizacion |
| Propiedades de datos | latitud y longitud | Propiedades que describen entidades |
| Instancias | 2483 | Cada instancia comprende un topónimo diferente |
| Axiomas | cerca(A,B):= cross(A,B) ∨ inside(A,B,C) ∨ touch(A,B) | Trata el concepto de ambigüedad espacial, donde: A,B,C son instancias (topónimos) |
| Gestor ontológico | Protégé V3.3 | Cuenta con el manejo de instancias sobre las clases |
| Editor para validación | RacerPorter V2.0 | Verificación de consistencia de la ontología |

Las clases constituyen las unidades básicas de la ontología que se pretende formalizar, a continuación se describen las clases de alto nivel de la ontología espacial:

- *Extructura_Geografica*: representa la clasificación más habitual de áreas o espacios artificiales creados por el ser humano. Cuenta con una subclase: *Construccion_o_Estructura_Humana que a su vez tiene diez subclases: Aeropuerto, Calle_o_Carretera, Camino, Canal, Ferrocarril, Monumento, Parque_Natural, Presa, Puente* y *Puerto.*

- *Extension_Geografica*: representa los espacios geográficos naturales; dentro de esta clase se incluyen las subclases: *Entidad_Geofisica*, que define los espacios naturales

marítimos y terrestres de México y la *Entidad_Geopolitica*, que representa la división política de México.

- *Localizacion*: representa una manera de definir la localización de un lugar.


**Figura 4.2. Taxonomía de la Ontología Espacial.**

**Relaciones**

Las relaciones de la ontología representan las interacciones y los enlaces entre los conceptos de dominio.

- *Propiedades de objeto*
  A partir de la jerarquía de clases presentada en la figura 4.2, se definen una serie de propiedades de objeto para, principalmente relacionar todas las clases de la ontología con la entidad *Localizacion*. A partir de las relaciones, es posible definir todos los aspectos que se quieren relacionar con la localización de los topónimos. A continuación en la Tabla 4.2 se detallan las relaciones descritas como propiedades de objeto incluidas en la ontología.

- *Propiedades de datos*
  Además de las propiedades que relacionan las diferentes entidades de la ontología, es necesario crear propiedades de datos que describen dichas entidades. Las principales en esta ontología son las que describen la entidad *Localizacion* y son las

siguientes *latitud* y *longitud*. Estas son consideradas propiedades funcionales ya que describen los valores de la localización de un topónimo.

**Tabla 4.2. Propiedades de objeto de la Ontología Espacial.**

| Propiedades de Objeto | Tipo | Descripción |
|---|---|---|
| contieneLocalizacion | Transitiva<br><br>A,B,C : localizaciones<br>Si $A \subset B \wedge B \subset C$,<br>$\Rightarrow A \subset C$ | Relaciona dos localizaciones para indicar que una localización puede contener a otra de longitud más reducida.<br>Propiedad inversa de *formaParteDeLocalización.* |
| formaParteDeLocalizacion | Transitiva<br><br>A,B,C : localizaciones<br>Si $A \subset B \wedge B \subset C$,<br>$\Rightarrow A \subset C$ | Relaciona dos localizaciones para indicar que puede formar parte de otra de mayor longitud.<br>Propiedad inversa de *contieneLocalización.* |
| tienePuntoDeInicio | Funcional<br><br>l: localización<br>A: punto inicial<br>l = l(A) | Relaciona una localización con su latitud y longitud, inicial.<br>Subpropiedad de *formaParteDeLocalización.* |
| tienePuntoFinal | Funcional<br><br>l: localización<br>B: punto final<br>l = l(A) | Relaciona una localización con su latitud y longitud del punto de finalización.<br>Subpropiedad de *formaParteDeLocalización.* |
| estaADistancia | Simétrica<br><br>A,B : puntos<br>d: distancia<br>d(A,B) = d(B,A) | Relaciona dos localizaciones para el cálculo de la distancia existente entre ellas. |

**Instancias**
Para comprobar la utilidad de la ontología, se incluyeron instancias que verifican el correcto funcionamiento de la ontología espacial, la cual está compuesta por 2483 instancias, donde cada una comprende a un topónimo diferente.

Para la tarea de desambiguación de topónimos se está utilizando como contexto un Corpus de noticias multilingüe (incluye los idiomas: Español, Inglés, Francés, Italiano y Portugués) de la tarea de búsqueda de respuestas (QA) de la iniciativa CLEF. Para la tarea que se está realizando sólo se ocupó el corpus en el idioma Español, el cual consta de 731 documentos. El corpus anterior comprende 216102 noticias de todo el mundo y cada una contiene topónimos.

La ontología espacial cubre 38% de noticias del total mencionado anteriormente, además de un 24.12% de topónimos diferentes contenidos en esas noticias. Esto es con respecto a las noticias de todo el mundo, así que se prosiguió a separar las noticias de México teniendo ahora un subconjunto del corpus y un total de 6633 noticias con contenidos relacionados con México. Ahora la ontología espacial abarca entonces un 99% de las noticias de México representando un 18.85% de topónimos diferentes que aparecen en esas noticias. Con estos resultados, nos damos cuenta que se cubre la mayor parte del subconjunto del corpus de noticias de México y de esta forma se verifica que la ontología espacial tiene un buen funcionamiento y considera a la mayor parte de los topónimos.

**Axiomas**
Se tiene como principal axioma de la ontología espacial la relación espacial "*cerca*", esta relación se diseñó y se incorporó para seguir verificando la consistencia de la ontología. Este axioma trata el concepto de ambigüedad espacial y se programó con las relaciones básicas de la ontología que son: *is_part_of* e *is_a*, teniendo como base la siguiente expresión:

$$\text{Si } \text{“A”} \wedge \text{“B”} \text{ is\_part\_of } \text{“C”} \quad \text{“A} \wedge \text{B” están cerca.}$$

Donde: A, B, C son topónimos de la Ontología produce el siguiente axioma:

$$\text{cerca}(A,B) := \text{cross}(A,B) \vee \text{inside}(\{A,B\}, C) \vee \text{touch}(A,B)$$

Desde el punto de vista especial, la relación "cerca", podría ser expresada en términos de predicados topológicos, tales como: *cross*, *inside*, y *touch*.

### 4.2.4. Validación de la Ontología Espacial

Para garantizar la consistencia de ontología espacial, ésta se validó utilizando un razonador espacial. En este proceso de validación, la información acerca de los objetos en el espacio y sus interrelaciones son recogidas por varios medios, tales como medidas, observaciones, o inferencia, y se utilizan para llegar a conclusiones válidas conforme a las relaciones de

objeto o para determinar la forma de realizar una tarea. El razonamiento espacial es usado para inferir todas las relaciones posibles entre un conjunto de objetos usando un subconjunto de las relaciones especificadas. RacerPro es un razonador utilizado tanto para Lógica Descriptiva Básica, como para muy expresiva y espacial, por este motivo Geontomex fue validada mediante este razonador. Además, también puede ser usado como un sistema para gestionar las ontologías de la Web Semántica basadas sobre OWL, es decir, puede ser usado como un motor para editores ontológicos como Protégé.

El editor de RacerPro utilizado para la validación, fue RacerPorter en el cual se pueden cargar bases de conocimiento, conmutar entre diferentes taxonomías, inspeccionar las instancias, visualizar TBoxes y ABoxes, manipular los servicios, entre otros servicios que la interfaz maneja, para la validación, razonamiento espacial y verificación de consistencia de ontologías.

### 4.2.5. Etiquetador Ontológico

Una vez creada la ontología que servirá como repositorio de sentidos para la tarea de desambiguación de topónimos, se prosiguió a crear un etiquetador ontológico que tiene como objetivo el identificar en el código creado por el editor Protégé (en este caso, código owl) cada uno de los componentes de la ontología, estos componentes son:

- Elementos ontológicos: Clases, Subclases, Propiedades, Instancias
- Etiquetas
- Comentarios

Los componentes identificados ayudarán a detectar los topónimos candidatos en el contexto, es decir, en el corpus. La tabla 4.3, muestra una parte de un ejemplo al ejecutar el etiquetador ontológico; este ejemplo es tomado del código owl que se generó al realizar la ontología espacial.

**Tabla 4.3. Parte de un ejemplo del resultado que da el etiquetador ontológico.**

| Componente Ontológico | Etiqueta Ontológica |
|---|---|
| Pais | CLASS |
| Continente | CLASS |
| formaParteDeLocalizacion | OBJECTPROPERTY |
| Nombre | DATATYPEPROPERTY |
| Puebla | INSTANCE |
| Ferrocarril | CLASS |
| El ferrocarril es un sistema de transporte…… | COMMENT |
| Entidad_Maritima | CLASS |
| http://www.w3.org/2001/XMLSchema#string | COMMENT |
| Bahia | CLASS |

## 4.3. Descripción del corpus

En esta sección se describe el corpus que sirve como contexto en la metodología para una de las tareas importantes de la recuperación de información geográfica como lo es la desambiguación de topónimos para el idioma Español. Esta etapa de la metodología está compuesta de cuatro componentes: el etiquetado del corpus con el etiquetador Tree Tagger, se extraen los términos multipalabra mediante *n_gramas*, se detectan esos términos en el corpus con la ayuda de la ontología y se hace un etiquetamiento manual para de esta forma tener un corpus etiquetado manualmente.

### 4.3.1. Corpus

Un corpus es un conjunto de textos recopilados, ya sea de un mismo tema o varios. El propósito de un corpus es convertirse en un conjunto de datos para proveer ejemplos de oraciones y ejemplos de uso de varias palabras para ser utilizados en algoritmos de aprendizaje automático. Dependiendo de la naturaleza de los algoritmos y de la tarea, las palabras que hay en un corpus puede estar previamente desambiguadas o no. Los corpora, es decir, colecciones de textos utilizados en los modelos de aprendizaje de idiomas; pueden tener un sentido anotado o crudo (es decir, sin etiquetamiento). Ambos tipos de recursos son útiles en los métodos de clasificación supervisados y no supervisados.

Existen diferentes definiciones o interpretaciones acerca de lo que es un corpus, una de ellas es la del Diccionario de la Real Academia Española (2012), la cual afirma que:

"*Un corpus es un conjunto ordenado y lo más extenso posible de datos o textos científicos, literarios, entre otros, que pueden servir de base en una investigación*".

Para poder explicar sus orígenes, tomaremos en cuenta la lingüística de corpus. Esta rama de la lingüística, es una metodología que se encarga de estudiar las diferentes leguas existentes tomando como base una gran cantidad de textos con datos reales, a los que se les denomina precisamente *corpus.* Durante la época conocida como *Early Corpus Linguistcs*, que abarcaba desde finales del siglo XIX hasta la década de los 50, los textos escritos y hablados eran considerados como fuentes primarias y únicas para las investigaciones científicas. Se puede decir que eran ciertamente estudios con índole empírico y estaban relacionados a diversas áreas de la lingüística como la adquisición del lenguaje o la propia enseñanza de la lengua, o inclusive, a estudios relacionados con los dialectos.

Si bien el corpus representa una herramienta muy importante para una investigación, cabe mencionar que como tal, tiene ciertas limitaciones, como las siguientes:

- Limitación temporal del corpus
  Se refiere a que un corpus está determinado por los limites de las muestras recopiladas y por tanto no puede abarcar la totalidad de un sistema que siempre es cambiante y por tanto, que está en constante evolución.

- Limitaciones por el volumen del corpus
  Un corpus puede ser de dimensiones enormes y por tanto contener un sinfín de datos, lo que como consecuencia dificultará su análisis y manejo.

Teniendo en cuenta criterios como: la modalidad de la lengua, el número de lenguas a que pertenecen los textos, el tamaño o cantidad de textos que conforman el corpus, los límites del corpus, la variedad lingüística o el grado de especialización de los textos, el periodo temporal que abarcan los textos, el tratamiento aplicado al corpus, entre otros. Se puede establecer la siguiente tipología:

a. Según la modalidad de la lengua
   - Corpus escritos: Los corpus textuales o escritos están conformados exclusivamente por muestras de lengua escrita.
   - Corpus orales: Los corpus orales, por su parte, únicamente recogen muestras de lengua hablada.
   - Corpus mixtos: Combinan ambas modalidades de lengua, aunque siempre favoreciendo la lengua escrita, ya que su obtención es menos costosa que la de la lengua oral que, además, requiere un proceso posterior de transcripción de las grabaciones.

b. Según el número de lenguas
   - Corpus monolingües: Están compuestos por textos en una sola lengua. Se recopilan con el objetivo de dar cuenta de dicha lengua o variedad lingüística (o de un subconjunto de la misma).
   - Corpus bilingües o multilingües: Están formados por textos de dos (bilingües) o más lenguas (multilingües) sin que, en principio, sean traducciones unos de otros y sin compartir criterios de selección.
   - Corpus alineados: Son corpus paralelos en los que, para facilitar su explotación, los textos están dispuestos unos al lado de otros por párrafos o frases, de tal forma que sea más fácil extraer las equivalencias de traducción: aquellos elementos que son traducciones mutuas.

c. Según la cantidad, la proporción y la distribución de los tipos de textos
   - Corpus grandes: No tienen un límite de palabras o este es muy elevado en comparación con otros tipos de corpus; no suelen atender a cuestiones de equilibrio o de representatividad. Cada vez es mayor la tendencia al aumento de volumen gracias a los medios y facilidades técnicas disponibles; no obstante, en

la actualidad existen corpus de gran tamaño diseñados con criterios que garantizan la representatividad de los datos.
- Corpus equilibrados: Recogen la misma proporción de diferentes tipos de textos.
- Corpus piramidales: Contienen textos distribuidos en estratos o niveles, de tal forma que un nivel consta de pocas variedades temáticas pero con muchos textos para cada una; un segundo nivel, de textos más variados temáticamente, pero con menos cantidad de cada uno; entre otros.
- Corpus léxicos ("*sample corpus*"): Recogen fragmentos de textos muy pequeños y de longitud constante en cada documento. Era lo habitual en los primeros corpus, debido a las limitaciones de tamaño que los medios técnicos de la época imponían.

d. Según los límites establecidos
   - Corpus cerrados: Constan de un número finito de palabras, que se establece de forma previa a la recopilación del corpus. Una vez alcanzado ese número, el corpus se da por finalizado, sin añadir más material posteriormente.
   - Corpus abiertos o corpus monitor: Son corpus dinámicos que se mantienen en constante crecimiento, normalmente mediante la introducción periódica de nuevas cantidades de textos según unas proporciones previamente definidas.

e. Según la especificidad de los textos
   - Corpus generales o de referencia: Pretenden reflejar la lengua o variedad lingüística de la forma más equilibrada posible; cuantos más tipos de textos, modalidades (textos orales, textos escritos), géneros y materias, mejor.
   - Corpus especializados: Recogen textos que puedan aportar datos para la descripción de un tipo particular de lengua ("sublenguaje").
   - Corpus genéricos: Recogen textos pertenecientes a un único género, ya que el objetivo es caracterizar ese género frente a otros.
   - Corpus canónicos: Están formados por todos los textos que configuran la obra completa de un autor.

f. Según el periodo temporal que abarcan los textos
   - Corpus periódicos o cronológicos: Recogen textos de unos años determinados o de unas épocas concretas con el objeto de estudiar la lengua producida durante ese período.
   - Corpus diacrónicos o históricos: Incluyen textos de diferentes etapas temporales sucesivas con el fin de poder observar evoluciones de la lengua en un período largo, lo que los diferencia de los corpus monitor, que no abarcan períodos temporales tan amplios.
   - Corpus sincrónicos: Su finalidad es permitir el estudio de una o más variedades lingüísticas en el momento presente, sin prestar atención a su evolución excepto en lo que se refiere a los cambios rápidos que ocurren en la actualidad.

g. Según el proceso al que se someta el corpus
   - Corpus simples, en bruto, no anotados o no codificados: Consisten en textos guardados sin formato alguno y sin añadir ningún tipo de información adicional, como pueden ser códigos o anotaciones.
   - Corpus verticales: Son el resultado de disponer en forma de columna las palabras de un texto ordenadas según criterios alfabéticos o de frecuencia. Las palabras se consideran aisladamente, sin contexto.
   - Corpus codificados o anotados: Están formados por textos a los que se les han añadido, de forma manual o automática, determinadas informaciones. Estas pueden ser codificaciones o anotaciones.

### 4.3.2. Descripción del Corpus

El corpus fue extraído de los textos escritos en la colección de noticias multilingüe de la tarea de búsqueda de respuestas (QA) de la iniciativa CLEF[32]. Este corpus es un corpus multilingüe y consta de los siguientes idiomas: Español, Francés, Inglés, Italiano, Portugués. En la tabla 4.4 se puede observar la descripción general de los idiomas del corpus. La colección de documentos utilizada en la metodología presentada consta de relatos periodísticos ocurridos en los años 1995 a 1996 de la agencia de noticias, para la metodología propuesta en esta tesis, se ocuparon únicamente los textos en Español.

**Tabla 4.4. Descripción general del corpus multilingüe.**

| Idioma | Colección | Período | # Documentos |
|---|---|---|---|
| Español | EFE<br>EFE | 1994<br>1995 | 731 |
| Francés | Le Monde<br>Le Monde<br>French SDA<br>French SDA | 1994<br>1995<br>1994<br>1995 | 1050 |
| Inglés | Loa Angeles Times<br>Glasgow Herald | 1994<br>1995 | 679 |
| Italiano | La Stampa<br>Italian SDA<br>Italian SDA | 1994<br>1994<br>1995 | 1076 |
| Portugués | Público<br>Público | 1994<br>1995 | 728 |

[32] Revisar: http://www.clef-initiative.eu//

La colección de documentos en Español contiene noticias y eventos de cobertura nacional e internacional que representan una amplia variedad de regiones geográficas y localizaciones. Esta colección consta de un total de 731 documentos y fue compuesta con noticias de la agencia EFE (agencia internacional de información o noticias, es la primera agencia de noticias multimedia en Español). Además proporciona colecciones en los idiomas: francés, inglés, italiano, portugués. Todas estas colecciones tienen una estructura común: información específica de periódico como fecha, página, tema, título, autor y el texto de la noticia. Las noticias están dadas en código *xml*, en la figura 4.3 se puede observar el formato de un fragmento de una noticia del corpus utilizado, las líneas de color café corresponden a topónimos del tipo GEO/GEO y las líneas de color verde corresponden a topónimos del tipo GEO/NO-GEO.

```
<CLAVE>FP0658</CLAVE>
<NUM>193</NUM>
<PRIORIDAD>U</PRIORIDAD>
<TITLE>   CICLISMO-RUTA MEXICO 94
           MEXICANO GONZALEZ ALCANZO LIDERATO AL GANAR SEGUNDA ETAPA

</TITLE>
<TEXT>    Zacatecas (México), 17 ene (EFE).- El ciclista mexicano Juan Luis
González, del equipo Zacatecas, alcanzó hoy el liderato de la Ruta
México'94 al ganar la segunda etapa, disputada entr el Puerto Madero
y Zacatcas, capital del estado del mismo nombre y situada al norte
del país.
   González protagonizó una escapada desde el principio de la etapa y
sólo tuvo la oposición de su compatriota Francisco de la Fuente, del
equipo Guanajuato, quien llegó a la meta poco después.
   Juan Luis González recorrió los 138 kilómetros de la etapa en tres
horas, 19 minutos y 24 segundos, mientras que De la Fuente llegó un
minuto y 17 segundos más tarde.
   Ambos ciclistas lograron aventajar en más de 13:30 minutos al
resto de competidores, que llegaron encabezados por el alemán Juergen
Werner, del equipo Telekom.
   De este modo, González arrebató el distintivo de líder de la
carrera al italiano Endrio Leoni, quien ganó la primera etapa,
disputada ayer en un circuito de 80 kilómetros.
   Mañana, martes, los 180 participantes partirán de Zacatecas hacia
la ciudad de San Luis Potosí, en una etapa que contará con 190
kilómetros de recorrido. EFE
   agm/sab
   01/18/01-41/94
```

**Figura 4.3. Formato de un fragmento de una noticia en Español.**

Una vez que se conoce el formato de las noticias se prosiguió a aplicar técnicas de preprocesamiento de textos, como puede ser el borrado de palabras sin significado (términos empleados y con poca utilidad) que producirán que algunas de la palabras no formen parte de la noticia. Es decir, limpiar el corpus de manera que solo se tenga la información que se desea ocupar. En este caso, la información requerida es el título de la noticia y el contenido de la misma; para este fin se hizo necesaria la creación de un programa que preprocesara cada documento, el programa fue realizado en *awk* que es

lenguaje de programación diseñado para procesar datos basados en texto, ya sean archivos o flujos de datos. Este programa dio como resultado la información necesaria para identificar los topónimos candidatos en las noticias, en la figura 4.4 se muestra la salida del programa realizado, esta salida consistió de un identificador que ayudará con a la indización del topónimo y el título seguido del contenido de la noticia. Este es el corpus inicial que se ocupó y de los731 documentos que se tenían, al ser preprocesados se obtuvo un total de 216102 noticias en Español de todo el mundo y cada una contiene topónimos.

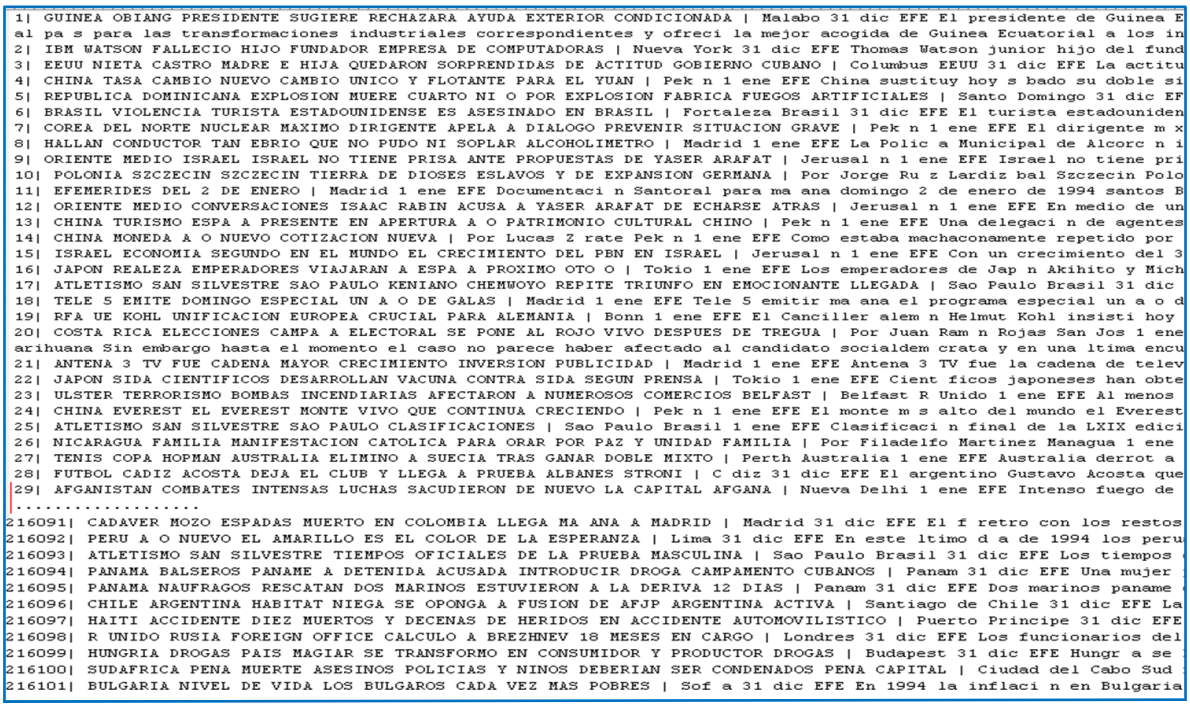


**Figura 4.4. Corpus inicial preprocesado.**

Una vez obtenido el corpus inicial se prosiguió con los siguientes componentes, en la siguiente sección se describirá el etiquetado del corpus y la detección de los topónimos en el corpus.

### 4.3.3. Etiquetamiento del Corpus

Etiquetar un corpus consiste en identificar los elementos léxicos simples y/o compuestos que lo integran, especificando las formas canónicas (lemas) a las que están asociados, su clase distribucional y las propiedades morfológicas de reflexión, en el caso de los verbos, los nombres, los adjetivos, los participios, entre otros. En la metodología presentada, el etiquetamiento del corpus se lleva a cabo mediante una herramienta para anotar textos con la información de la parte de su discurso (*part-of-speech*, *POS*) y del lema, esta herramienta se llama TreeTagger.

TreeTagger[33] es un lenguaje independiente que etiqueta las partes del discurso de una oración, ha sido utilizado con éxito para etiquetar en los idiomas: Alemán, Inglés, Francés, Italiano, Holandés, Español, Búlgaro, Ruso, Griego, Portugués, Chino, Swahili, Latín, Estonio y viejos textos en Francés, es adaptable a otros idiomas si están disponibles un diccionario léxico y un corpus de entrenamiento etiquetado manualmente.

Los parámetros del etiquetador TreeTagger utilizados para el etiquetamiento POS en la metodología fueron los archivos en Español y los archivos en Español con codificación UTF-8; se utilizaron ambos parámetros para observar cual de ellos tenía mejor resultado de etiquetamiento en el corpus. Es decir, se hizo una revisión aleatoria manual de algunos de los resultados que dio cada archivo, esta revisión consistió en observar la etiqueta y el lema de las palabras en cada resultado de salida, estas palabras fueron las mismas en cada revisión. Se consideró como mejor resultado el archivo de salida que tenía mas palabras etiquetadas y con su lema correcto con base en esta revisión se concluyó que los parámetros en Español con codificación UTF-8 dieron un mejor resultado y por tal motivo fue seleccionado para etiquetar el corpus y fue el que se prosiguió a ocupar a lo largo de la metodología. En la figura 4.5 se muestra un ejemplo de salida del etiquetador, este ejemplo es una parte del corpus utilizado. La salida del resultado del etiquetado consiste de tres columnas, la primera es la palabra, seguida de la parte de la oración (etiquetado POS) y el lema.

| Palabra | POS | Lema |
|---|---|---|
| La | ART | el |
| precipitación | NC | <unknown> |
| de | PREP | de |
| lluvia | NC | lluvia |
| en | PREP | en |
| la | ART | el |
| ciudad | NC | ciudad |
| de | PREP | de |
| Tuxtla | NC | <unknown> |
| Gutiérrez | NP | Gutiérrez |
| del | PDEL | del |
| estado | NC | estado |
| de | PREP | de |
| Chiapas | NP | Chiapas |
| alcanzó | VLfin | alcanzar |
| los | ART | el |
| 50 | CARD | @card@ |
| cm3 | NC | <unknown> |
| el | ART | el |
| día | NC | día |
| de | PREP | de |
| ayer | ADV | ayer |
| , | CM | , |
| esto | DM | este |
| provocó | VLfin | provocar |
| un | ART | un |
| gran | ADJ | grande |
| desastre | NC | desastre |
| vial | ADJ | vial |
| . | FS | . |

**Figura 4.5. Ejemplo de salida del etiquetador TreeTagger.**

[32] Disponible en: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Para tener una mejor comprensión de la figura 4.5, se sugiere consultar el Anexo A que se encuentra al final de esta tesis, que contiene las etiquetas POS y su correspondiente descripción.

El etiquetamiento fue de gran ayuda debido a que el corpus no estructurado y sin identificación de topónimos candidatos, se etiqueta morfológicamente cada noticia; con la finalidad de incluir las categorías morfológicas como categorías dentro de un clasificador supervisado. Además de que facilita el etiquetamiento manual.

El etiquetado de los componentes del texto consiste en etiquetar *tokens* o elementos invisibles del texto secuencialmente con etiquetas sintácticas, como nos hemos dado cuenta Por otra parte, el reconocimiento de entidades con nombre se ocupa de la búsqueda en el texto de referencias que pertenecen a un conjunto definido de categorías, como nombres de personas, nombres de organizaciones o nombres de lugares, entre otros. En esta metodología se ha hecho la combinación de ambas técnicas presentando resultados favorables para la obtención de los candidatos a posibles topónimos. Además de que se determinan topónimos candidatos en las noticias usando la ontología espacial antes descrita, que describe el espacio geográfico considerando los objetos geográficos naturales y artificiales de la República Mexicana.

### 4.3.4. Detección de topónimos en el Corpus

En la figura 4.6 se muestra el diagrama de detección del topónimo en el corpus, este proceso consistió en utilizar el corpus y la ontología, con estos dos recursos se detectan los topónimos en el corpus. Con el corpus se extraen los términos que lo componen y con la ontología se comparan los topónimos para así ser detectados en el corpus.

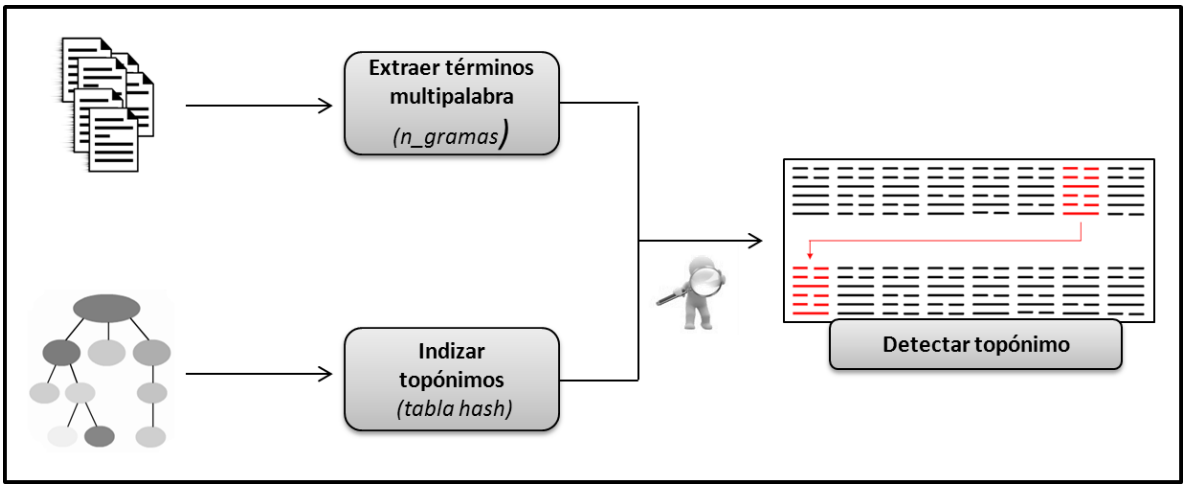


**Figura 4.6. Diagrama de detección de un topónimo en el corpus.**

Este proceso consistió en extraer términos multipalabra (*n_gramas*) en el corpus, es decir, identificar los *n_gramas* o secuencias de elementos, en este caso palabras, que están en el corpus. Se identificaron desde un hasta siete *n_gramas*, por que existen topónimos que incluyen sólo una palabra y el topónimo de máxima longitud que instancia la ontología contiene siete palabras. Este topónimo es "La Unión de Isidoro Montes de Oca", es decir, por lo tanto los valores para *n* en los *n_gramas* fueron 1, 2, 3, 4, 5, 6 y 7.

Por otro lado, se tiene la ontología espacial que se construyó y esta es indizada por medio de una tabla hash, debido a que el acceso mediante esta para detectar el topónimo es más rápido. Una de las principales características y operación que soporta una tabla hash de manera eficiente es la búsqueda, ya que permite el acceso a los elementos almacenados a partir de una clave generada. Para el desarrollo de la metodología la clave generada fue un número consecutivo, debido a que lo que tenía mayor importancia es el detectar el topónimo de manera más rápida, no tanto el índice que la tabla hash generaba. Entonces se identifican todos aquellos términos en las noticias (corpus) que sean calificados como topónimos en la ontología, es decir, que sean instancias de la ontología.

Una vez que se han extraído los *n_gramas* y se ha indizado la ontología, se prosigue a identificar los topónimos en el corpus, esta identificación o búsqueda se hace empatando el *n_grama* extraído con la instancia indizada de la ontología. Cuando se hace el empatamiento, los términos son seleccionados y guardados en un archivo, además de que las noticias que contienen estos términos son guardados en otro archivo con el fin de guardar la información necesaria que servirá como características para el modelo de clasificación.

La búsqueda en la ontología, además de comprobar si un nombre de lugar candidato es una verdadera referencia geográfica, proporciona la desambiguación con base a la profundidad en la que se encuentran los posibles referentes.

A la hora de etiquetar el corpus e identificar entidades con nombres tal como se mencionó en la sección anterior (sección 4.3.3) que se obtuvieron resultados favorables al combinar ambas técnicas a la hora de obtener los candidatos a posibles topónimos. Sin embargo, a pesar de esos resultados favorables, al traducir esos candidatos a verdaderas referencias geográficas es donde aparecen los problemas relacionados con la ambigüedad. Se puede decir que este es un problema a medio camino entre las dos etapas que componen la geo-referenciación. En primer lugar, hay que determinar si los candidatos son verdaderas referencias geográficas o no y en segundo lugar desambiguar esas referencias. En la siguiente sección (sección 4.4) se describe que una de las etapas de la metodología propuesta es precisamente el método desambiguador y este está basado en un modelo de clasificación.

## 4.4. Método desambiguador: Modelo de clasificación

En esta sección se describe el método desambiguador que está basado en un modelo de clasificación para la metodología de desambiguación de topónimos propuesta. Esta etapa se compone del método de enriquecimiento de corpus a partir de la Web y del modelo de clasificación para desambiguar los topónimos. En el modelo de clasificación se utilizaron métodos supervisados como árboles de clasificación, Naïve Bayes y Máquina de Soporte Vectorial (SVM).

### 4.4.1. Método de enriquecimiento de corpus

Dado que el objetivo final del trabajo de investigación es la desambiguación de topónimos mediante técnicas de categorización supervisada (clasificadores probabilísticos o máquinas de soporte vectorial), se requiere esencialmente de un corpus supervisado que permita obtener un modelo adecuado de clasificación. Hasta donde se sabe, tal corpus no se encuentra disponible para la tarea de desambiguación de topónimos tipo GE/NO-GEO. Por tanto se hizo uso de la técnica propuesta en [8] para crear un corpus de entrenamiento usando técnicas de enriquecimiento de colecciones de textos. Este corpus de entrenamiento (supervisado) podrá ser entonces usado en algunos de los clasificadores que se ocuparon como árboles de clasificación, Naïve Bayes y máquina de soporte vectorial.

Los pasos para la creación del corpus supervisado se muestran en la figura 4.7. Básicamente se trata de una técnica de *bootstrapping*, en donde se crea un conjunto pequeño de textos clasificados manualmente (corpus inicial) que es enriquecido mediante un proceso de búsqueda de oraciones en la Web y su posterior evaluación para garantizar la calidad de la oración para el corpus de entrenamiento. Las oraciones extraídas de la Web son lo que comúnmente conocemos como *snippets*, que son una descripción resumida de un sitio, que extraen los motores (máquinas) de búsqueda Web cuando se hace una consulta, y que son mostrados en los resultados -a modo de resumen- junto a la URL.

En el caso que nos atañe, es decir, la desambiguación de topónimos, la creación de un corpus inicial etiquetado manualmente significa que los textos (noticias) tienen una referencia GEO (muestras positivas) o tienen una referencia NO-GEO (muestras negativas). Dado que los referentes o topónimos GEO/NO-GEO son etiquetados manualmente en el corpus inicial, entonces, se puede confiar en la calidad de esta colección. Sin embargo, es importante garantizar que las muestras (tanto positivas como negativas) sean lo suficientemente representativas, como para poder generalizar el corpus inicial, enriqueciéndolo con muestras extraídas de la Web.

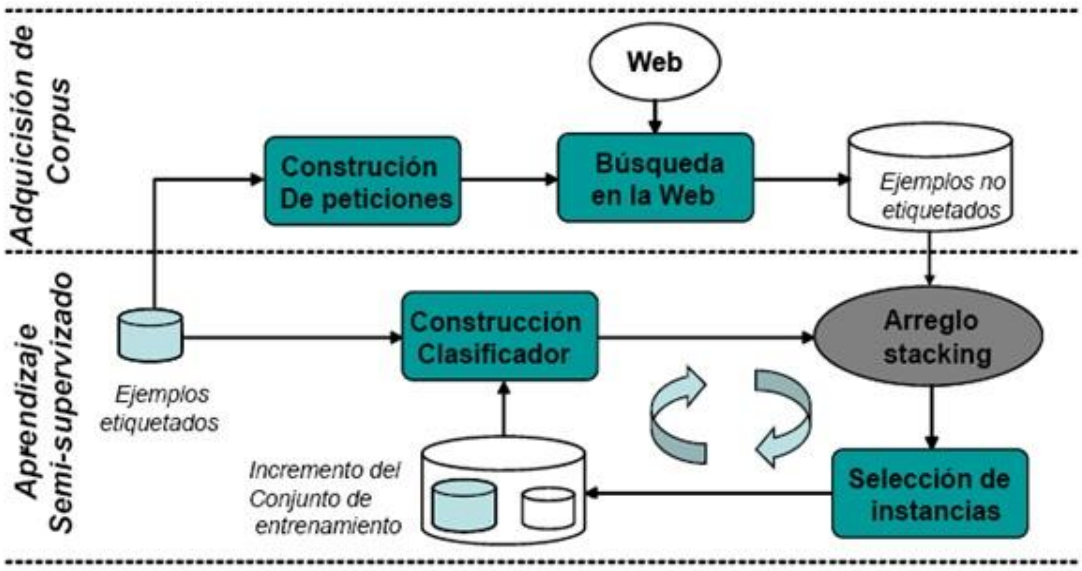


**Figura 4.7. Proceso de enriquecimiento de Corpus a partir de la Web [8].**

El corpus inicial fue extraído de los textos escritos en Español del corpus descrito en la Sección 4.3 y recordando se trata de una colección de noticias multilingüe de la tarea de búsqueda de respuestas de la iniciativa CLEF. Tal y como lo requiere la técnica de enriquecimiento, una vez constituido el corpus inicial, se procede a enriquecerlo. El procedimiento considera obtener muestras de la Web usando algún sistema de recuperación de información. En nuestro caso, usamos el API de Google y las consultas fueron creadas a partir de n-gramas (bigramas y trigramas) de palabras extraídas de la noticia etiquetada. Se consideraron únicamente aquellos *n_gramas* que incluyan al topónimo. Por ejemplo, para el siguiente fragmento de texto (perteneciente a una muestra positiva):

```
En la ciudad de Guadalajara hubo un enfrentamiento ...
```

Se obtienen los cinco *n_gramas* siguientes: {de Guadalajara}, {Guadalajara hubo}, {ciudad de Guadalajara}, {de Guadalajara hubo} y {Guadalajara hubo un}. Estos *n_gramas* se usan como consulta en Google y se lanzan las peticiones de búsqueda a la Web con la finalidad de descargar información que contenga algunos de los *n-gramas* usados como consulta.

El objetivo es encontrar más ejemplos que potencialmente tengan relación con el topónimo en cuestión. En este caso, hemos solicitado 500 *snippets* de Google por cada consulta, lo que arroja un total de 2,500 *snippets* asociados a un cierto topónimo. Es claro que no todos los *snippets* tienen la calidad para ser incluidos en el corpus de entrenamiento (enriquecimiento), y por tanto su calidad se evalúa usando un modelo de clasificación construido sobre las muestras positivas y negativas del corpus inicial. Es decir, si el *snippet* es clasificado como muestra positiva con un umbral mayor a 85%, entonces, éste se incluye en el corpus inicial, se reconstruye el modelo de clasificación, ahora considerando la nueva

muestra y se repite el proceso para un nuevo *snippet* extraído de la Web. El proceso se realiza tanto para muestras positivas como para muestras negativas. El corpus final de entrenamiento está constituido por 3,222 muestras positivas (tipo GEO) y 3,682 muestras negativas (tipo NO-GEO). Una vez construido el corpus de entrenamiento, entonces se puede proceder a realizar el proceso de desambiguación de topónimos.

Dado que no existe un corpus estándar para la evaluar la tarea de desambiguación de topónimos en Español, entonces usamos como *baseline* los resultados de la clasificación sin tener en cuenta el proceso de enriquecimiento del corpus inicial. En el siguiente capítulo se presentan los resultados obtenidos al aplicar la metodología propuesta.

### 4.4.2. Modelo de clasificación

El problema de clasificación es uno de los principales que aparecen en la actividad científica y constituye un proceso circunstancial con casi cualquier actividad humana. De tal manera que en la resolución de problemas en la toma de decisiones, la primera tarea consiste precisamente en clasificar el problema o la situación, para después aplicar la metodología correspondiente y que en buena medida dependerá de esa clasificación. En esta metodología para la desambiguación de topónimos se propone desambiguar mediante un modelo de clasificación, donde desambiguar es equivalente a *clasificar* a un topónimo tipo GEO/NO-GEO.

En la tarea de clasificación el conjunto de ejemplos a utilizar se divide en dos conjuntos independientes, uno es nombrado conjunto de entrenamiento y el segundo conjunto de evaluación, como se muestra en la figura 4.8. El proceso de clasificación se divide en dos fases las cuales se describen a continuación:

- Construcción del modelo de clasificación
  En este paso, se obtiene a través de un algoritmo de clasificación, un modelo que de acuerdo a los datos analizados, permite una clasificación de los mismos con una precisión global y por clases. Este resultado se obtiene a partir de una etapa conocida como entrenamiento (*training*), en el cual el algoritmo analiza los datos para obtener una representación del comportamiento de los mismos.

- Evaluación del modelo de clasificación
  En la segunda fase se realiza una evaluación sobre el comportamiento del modelo encontrado en la fase anterior utilizando el conjunto de pruebas (*testing*), el ejemplo lo podemos observar en la figura 4.8.

Cabe mencionar que los métodos basados en clasificación tienen como ventajas: poder ser aplicados a distintos tipos de variables predictivas ya sean continuas o categóricas, los resultados son fáciles de entender e interpretar, es rápido de calcular, no tiene problema al trabajar con datos perdidos y automáticamente realiza la selección de variables.

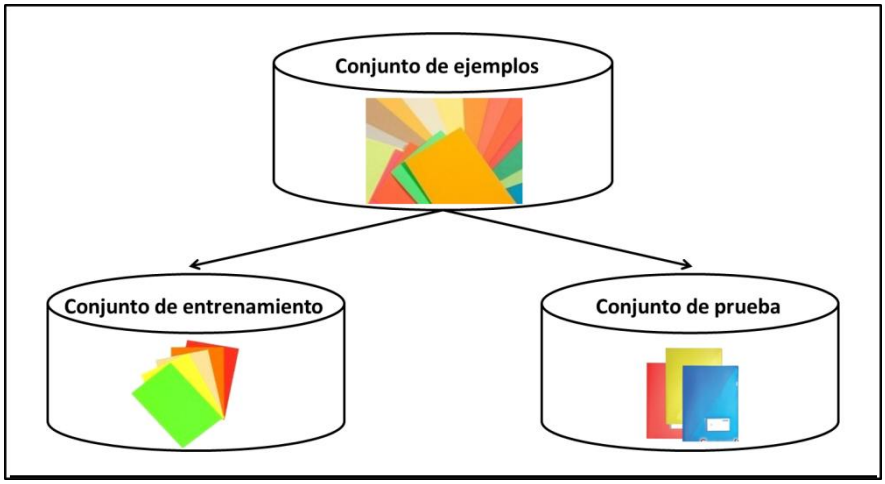


**Figura 4.8. División del conjunto de ejemplos en conjunto de entrenamiento y datos de prueba.**

Las ventajas anteriormente mencionadas hacen de los métodos de clasificación candidatos idóneos para situaciones en donde el objetivo sea buscar posibles tendencias o encontrar grupos de datos con características similares.

### 4.4.3. Métodos supervisados

En los últimos 15 años, la comunidad del PLN ha sido testigo de un cambio significativo del uso de sistemas manualmente diseñados para el empleo de métodos de clasificación automatizados. Este espectacular aumento de interés hacia las técnicas de aprendizaje se refleja por el número de enfoques supervisados aplicados a la desambiguación del sentido de las palabras. Debido a este hecho, en esta metodología se utilizaran técnicas de aprendizaje supervisado para la desambiguación de topónimos. Por lo general, realizar una tarea de clasificación se refiere a que a una oración (noticia) se le asigne la etiqueta (GEO/NO-GEO) apropiada a esa instancia de la oración.

El conjunto de entrenamiento usado para que un clasificador aprenda, típicamente contiene una serie de ejemplos (corpus inicial) en los que un topónimo dado es etiquetado manualmente. En general, los enfoques supervisados para la desambiguación han tenido mejores resultados que los enfoques no-supervisados [44]. A continuación se explicarán brevemente los métodos supervisados aplicados a esta metodología.

## Árboles de clasificación

Los métodos basados en árboles de clasificación tienen la virtud de generar modelos de descripción de los datos, los cuales pueden ser utilizados para separar los datos en clases a partir de identificar características comunes entre cada clase. Estos resultados pueden utilizarse en predicciones de futuras tendencias.

Un árbol de decisión o clasificación es un modelo predictivo usado para representar reglas de clasificación con una estructura de árbol que recursivamente divide el conjunto de datos de entrenamiento. Cada nodo interno de un árbol de decisión representa una prueba (*training*) en una función de valor y cada rama representa un resultado de la prueba (*testing*). Una predicción se realiza cuando se llega a un nodo terminal (es decir, una hoja). Un algoritmo popular para el aprendizaje de los árboles de decisión es el algoritmo C4.5, una extensión del algoritmo ID3.

El algoritmo J48 de Weka[33], fue utilizado para el modelo de clasificación para la tarea que nos atañe, la desambiguación de topónimos. El algoritmo es una implementación del algoritmo C4.5, uno de los algoritmos de minería de datos que más se ha utilizado en multitud de aplicaciones. En un experimento comparativo con las máquinas de varios algoritmos de aprendizaje para WSD, Mooney [69] llegó a la conclusión de que los árboles de decisión obtenidos con el algoritmo C4.5 son superados por otros métodos supervisados, con base a esto fue utilizado. De hecho, aunque representan un modelo predictivo de manera compacta y legible, sufren varios problemas, como la escasez de datos debido a las características con un amplio número de valores, la falta de fiabilidad de las predicciones por conjuntos de entrenamiento pequeños, entre otros. Un ejemplo de un árbol de decisión se muestra en la figura 4.9; supongamos que tenemos el topónimo del tipo GEO/NO-GEO "Benito Juárez", en la frase "Benito Juárez es uno de los diez municipios que integran el estado mexicano de Quintana Roo", el árbol es eficaz siguiendo el camino - si, si, si - , la elección del sentido (hoja) "Municipio" se hace.

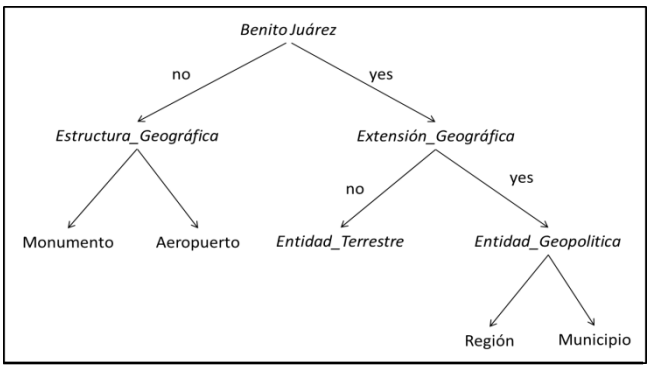


**Figura 4.9. Ejemplo de un árbol de clasificación para un topónimo.**

[33] Disponible en: http://www.cs.waikato.ac.nz/ml/weka/

**Naïve Bayes**

El clasificador Naïve Bayes es un clasificador probabilístico basado en el Teorema de Bayes. Se basa en el cálculo de probabilidad condicional de cada sentido de una palabra *w* dadas las características $f_j$ en el contexto. El sentido *S* que maximizan la fórmula (8) es elegido como el sentido más apropiado en el contexto.

$$
\begin{aligned}
S &= argmax_{S_i \in Senses_D(w)} P(S_i \mid f_{1,\ldots,} f_m) \\
&= argmax_{S_i \in Senses_D(w)} \frac{P(f_{1,\ldots,} f_m \mid S_i)\, P(S_i)}{P(f_{1,\ldots,} f_m)} P(S_i \mid f_{1,\ldots,} f_m) \\
&= argmax_{S_i \in Senses_D(w)} P(S_i) \prod_{i=1}^{m} P(f_j \mid S_i)
\end{aligned}
\tag{8}
$$

De la fórmula (8), tenemos que *m* es el número de características y la última fórmula se obtiene basándose en el ingenuo supuesto de que las características son condicionalmente independientes dado el sentido (el denominador se desecha, ya que no influye en los cálculos). Las probabilidades $P(S_i)$ y $P(f_j|S_i)$ se calculan respectivamente, como las frecuencias de ocurrencia relativa en el conjunto de entrenamiento del sentido $S_i$ y la presencia de la función $f_j$ en ese sentido. A pesar de la suposición de independencia, el método se compara bien con otros métodos supervisados.

Dentro de las ventajas de utilizar este clasificador probabilístico en la metodología propuesta tenemos las siguientes: el método es robusto al posible ruido presente en los ejemplos de entrenamiento y a la posibilidad de tener entre esos ejemplos de entrenamiento datos incompletos o posiblemente erróneos. El algoritmo de Naïve Bayes en la plataforma Weka está basado en redes neuronales.

Naïve Bayes es un método importante no sólo por que ofrece un análisis cualitativo de los atributos y valores que pueden intervenir en el problema, si no por que se da cuenta también de la importancia cuantitativa de esos atributos. En el aspecto cualitativo podemos representar cómo se relacionan esos atributos ya sea en una forma causal, o señalando simplemente de la correlación que existe entre esas variables (o atributos). Cuantitativamente (y ésta es la gran aportación de los métodos bayesianos), da una medida probabilística de la importancia de esas variables en el problema (y por lo tanto una probabilidad explícita de las hipótesis que se formulan). Esta es quizá una de las diferencias fundamentales que ofrecen las redes bayesianas con respecto a otros métodos - como puedan ser los árboles de decisión y las redes neuronales -, que no dan una medida cuantitativa de esa clasificación.

**Máquinas de Soporte Vectorial**

Este método introducido por Boser [70] se basa en la idea de un aprendizaje hiperplano lineal del conjunto de entrenamiento que separa ejemplos positivos de los ejemplos negativos. El hiperplano se encuentra en ese punto del hiperespacio que maximiza la distancia a los ejemplos positivos y negativos más cercanos (llamados vectores de soporte). En otras palabras, las máquinas de soporte vectorial (*Vector Support Machines*, SVM) tienden al mismo tiempo para minimizar el error de clasificación empírica y maximizar el margen geométrico entre los ejemplos positivos y negativos. En la figura 4.10 ilustra la intuición geométrica: la línea en negrita representa el plano que separa las dos clases de ejemplos, mientras que las dos líneas continuas representan el plano tangente a los ejemplos más cercanos positivos y negativos.

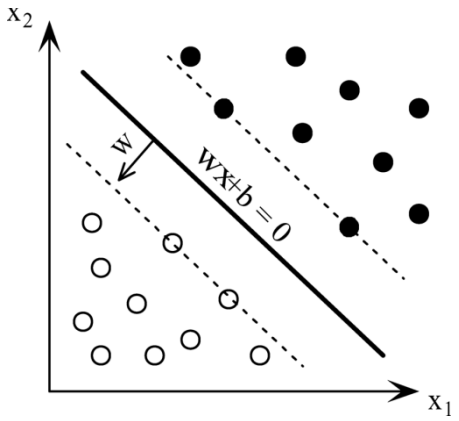


**Figura 4.10. Intuición geométrica de SVM.**

El clasificador lineal se basa en dos elementos: un vector de pesos *w* perpendicular a el hiperplano (que representa un conjunto de entrenamiento y componentes que representan características) y un sesgo *b* que determina el desplazamiento del hiperplano desde el origen. Un ejemplo *x* sin etiqueta se clasifica como positivo si $f(x) = w \cdot x + b \geq 0$ (negativo lo contrario).

Puede suceder que el hiperplano no puede dividir el espacio lineal. En este caso es posible utilizar variables de holgura para "ajustar" el conjunto de entrenamiento, y permitir una separación lineal del espacio.

Como SVM es un clasificador binario, puede ser adaptado y utilizado para el tipo de desambiguación que se está realizando, tenemos un clasificador binario, como resultado, el sentido con la más alta confianza se selecciona. SVM ha sido aplicado a números problemas de PLN, incluyendo categorización de textos, etiquetamientos, WSD, entre otros. SVM ha sido el clasificador que muestra los resultados más favorables para la tarea que se está resolviendo en comparación con árboles de clasificación y Naïve Bayes.

# CAPÍTULO 5. PRUEBAS Y RESULTADOS

En este capítulo, se muestran y discuten los resultados experimentales obtenidos al aplicar la metodología propuesta para la tarea de desambiguación de topónimos, el análisis se centra principalmente en el modelo de clasificación. Se observa que tras aplicar la metodología propuesta se obtienen resultados favorables.

# Capítulo 5

# Pruebas y resultados

En este capítulo, se presentan las pruebas y los resultados que se han obtenido al utilizar los siguientes métodos supervisados: árboles de clasificación (J48), Naïve Bayes (NB), Naïve Bayes Multinomial (NBM) y Máquinas de Soporte Vectorial (SVM); para el modelo de clasificación propuesto en la metodología general para la desambiguación de topónimos. En particular, los experimentos fueron realizados para el corpus inicial y el corpus enriquecido con la información descargada de la Web. Se destaca que los topónimos candidatos se recuperan si son instancias de la ontología realizada. El enriquecimiento de corpus se realiza utilizando técnicas de *bootstrapping*. Se proporciona un corpus supervisado validado manualmente y posteriormente se obtienen los *snippets* de la Web para incrementar el tamaño del corpus inicial.

A continuación se presentan los resultados de los experimentos que se realizaron con el modelo de clasificación para desambiguar topónimos. El corpus inicial con el cual se trabajó es el corpus de noticias descrito en la sección 4.3; las noticias habitualmente conllevan un alto contenido geográfico, por ejemplo, los geodatos extraídos representan el lugar donde se produjo la noticia. Posteriormente, este corpus se enriqueció con información descargada de la Web.

## 5.1. Resultados experimentales sobre el corpus inicial

En esta sección se muestran los resultados experimentales obtenidos al evaluar diferentes conjuntos de prueba sobre el corpus inicial, es decir, sin considerar el enriquecimiento del corpus.

En la Tabla 5.1.1 se muestran los valores de precisión y recuerdo (*recall*) para el primer conjunto (*test 1*) de prueba con respecto al corpus inicial. La Figura 5.1.1 muestra que el clasificador J48 es más preciso con respecto al resto de los clasificadores; sin embargo, su recuerdo es muy bajo. Por otro lado, el clasificador SVM obtiene el mejor recuerdo, pero también el peor valor de precisión. Cuando se calcula la media armónica entre precisión y

recuerdo (*F-Measure*), se observa que es el modelo de SVM el que obtiene el mejor balance entre estas dos medidas.

**Tabla 5.1.1. Valores de evaluación obtenidos por los diferentes clasificadores sobre el corpus inicial con el conjunto de prueba 1.**

| *Clasificador* | *Precisión* | *Recuerdo* | *F-Measure* |
|---|---|---|---|
| **J48** | 0.88 | 0.12 | 0.2112 |
| **NB** | 0.74 | 0.26 | 0.3848 |
| **NBM** | 0.72 | 0.28 | 0.4032 |
| **SVM** | 0.68 | 0.32 | 0.4352 |

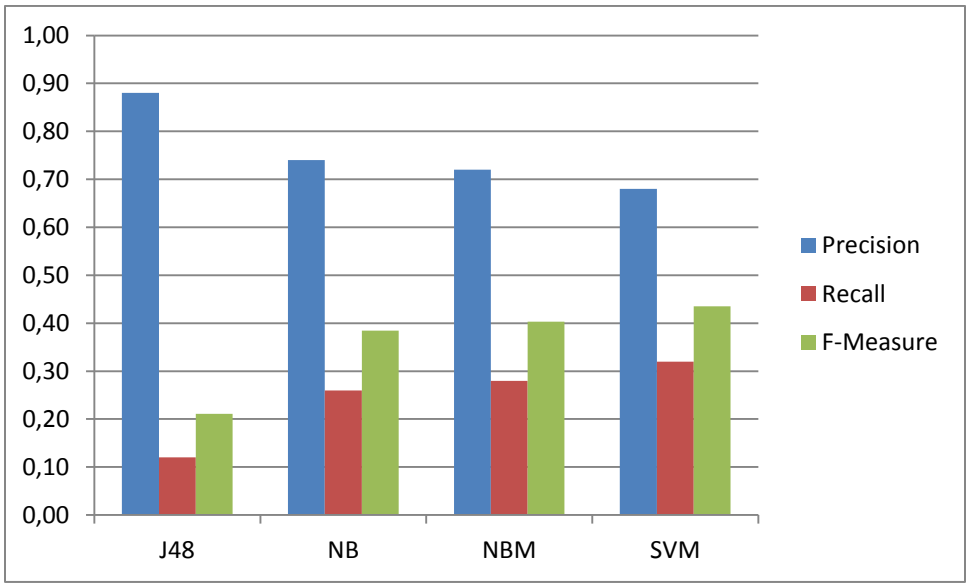


**Fig. 5.1.1. Evaluación de los clasificadores sobre el conjunto de prueba 1.**

Aun así, estos resultados no son concluyentes para determinar el mejor clasificador para esta tarea. Por tanto, a continuación se muestran los resultados obtenidos para otros conjuntos de prueba sobre el mismo corpus inicial.

En la Tabla 5.1.2 se muestran los valores de precisión y recuerdo para otro conjunto de prueba con respecto al corpus inicial, llamemos a este conjunto de prueba 2 (*test 3*). La Figura 5.1.2 muestra que el clasificador J48 es más preciso con respecto al resto de los clasificadores; sin embargo, su recuerdo es muy bajo. Por otro lado, NBM y SVM tienen un comportamiento uniforme, es decir, los valores de evaluación son similares y a pesar de este hecho y notando que NB obtiene el mejor recuerdo, pero también el peor valor de precisión. Cuando se calcula la media armónica entre precisión y recuerdo, se observa que es el modelo de NB el que obtiene el mejor balance entre estas dos medidas, pero se sigue observando que SVM mantiene un comportamiento estable con respecto al primer conjunto de prueba, es decir, el valor del *F-Measure* aumenta en este conjunto de prueba.

**Tabla 5.1.2. Valores de evaluación obtenidos por los diferentes clasificadores sobre el corpus inicial con el conjunto de prueba 2.**

| *Clasificador* | *Precisión* | *Recuerdo* | *F-Measure* |
|---|---|---|---|
| **J48** | 0.78 | 0.22 | 0.3432 |
| **NB** | 0.58 | 0.42 | 0.4872 |
| **NBM** | 0.64 | 0.36 | 0.4608 |
| **SVM** | 0.64 | 0.36 | 0.4608 |

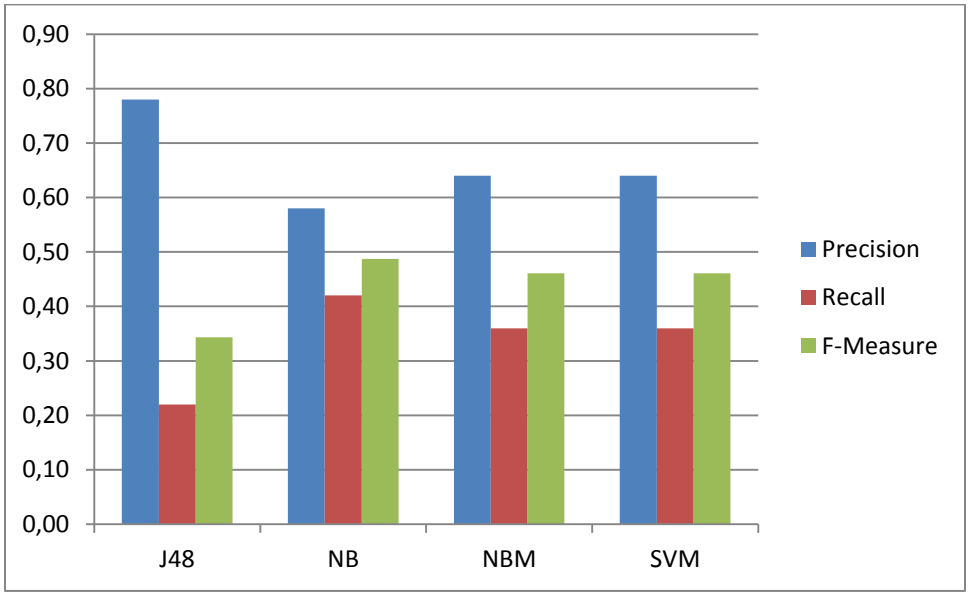


**Fig. 5.1.2. Evaluación de los clasificadores sobre el conjunto de prueba 2.**

Ahora observemos que sucede con un tercer conjunto de prueba (*test 3*), en la Tabla 5.1.3 se muestran los valores de precisión y recuerdo para otro conjunto de prueba con respecto al corpus inicial.

**Tabla 5.1.3. Valores de evaluación obtenidos por los diferentes clasificadores sobre el corpus inicial con el conjunto de prueba 3.**

| *Clasificador* | *Precisión* | *Recuerdo* | *F-Measure* |
|---|---|---|---|
| **J48** | 0.84 | 0.16 | 0.2688 |
| **NB** | 0.86 | 0.14 | 0.2408 |
| **NBM** | 0.86 | 0.14 | 0.2408 |
| **SVM** | 0.88 | 0.12 | 0.2112 |

La Figura 5.1.3 muestra que el clasificador SVM es más preciso con respecto al resto de los clasificadores; sin embargo, su recuerdo es bajo al igual que la media armónica entre precisión y recuerdo. Este hecho, permite observar que en este conjunto de prueba el clasificador SVM tiene un mejor comportamiento. Pero también se observa que NB y NBM tienen un comportamiento similar, esto es debido a que, NBM es un método popular para la clasificación de documentos debido a su eficiencia computacional y la predicción de rendimiento relativamente bueno y solo tiene variaciones con respecto a la fórmula de clasificación.

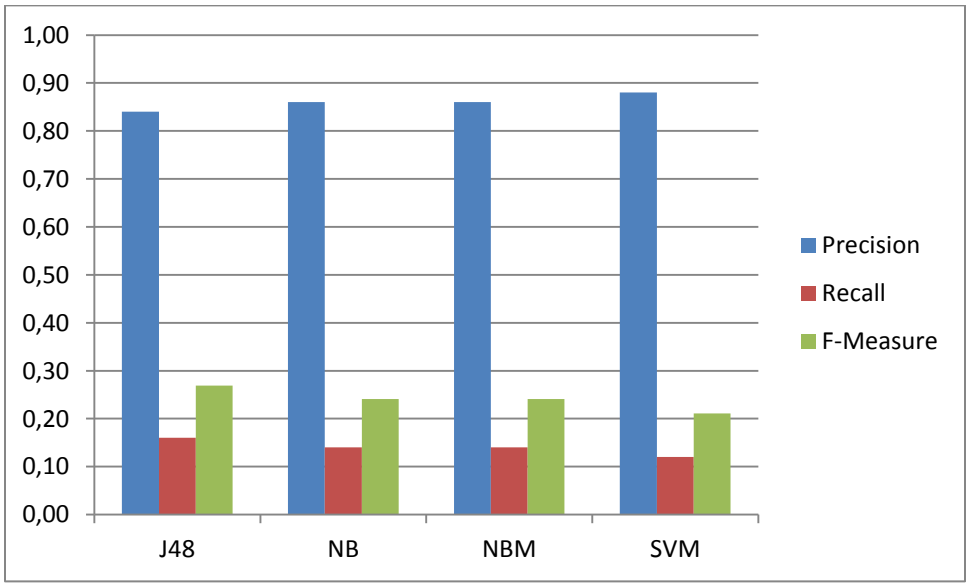


**Fig. 5.1.3. Evaluación de los clasificadores sobre el conjunto de prueba 3.**

En la Tabla 5.1.4 se muestran los valores de precisión y recuerdo para el último conjunto con respecto al corpus inicial, a este conjunto lo llamaremos conjunto de prueba 4 (*test 4*). La Figura 5.1.4 muestra que el clasificador NBM es más preciso con respecto al resto de los clasificadores. Sin embargo, su recuerdo es muy bajo. Por otro lado, el clasificador SVM obtiene un valor medio precisión y también de recuerdo, con respecto a los valores de los demás clasificadores. Además de que al calcular la media armónica entre precisión y recuerdo de este mismo clasificador, se observa que sigue existiendo un valor medio. Sin embargo, el valor de precisión para el clasificador J48 es el más bajo, pero obtiene el mejor balance entre precisión y recuerdo.

**Tabla 5.1.4. Valores de evaluación obtenidos por los diferentes clasificadores sobre el corpus inicial con el conjunto de prueba 4.**

| *Clasificador* | *Precisión* | *Recuerdo* | *F-Measure* |
|---|---|---|---|
| **J48** | 0.68 | 0.32 | 0.4352 |
| **NB** | 0.8 | 0.2 | 0.32 |
| **NBM** | 0.74 | 0.26 | 0.3848 |
| **SVM** | 0.72 | 0.28 | 0.4032 |

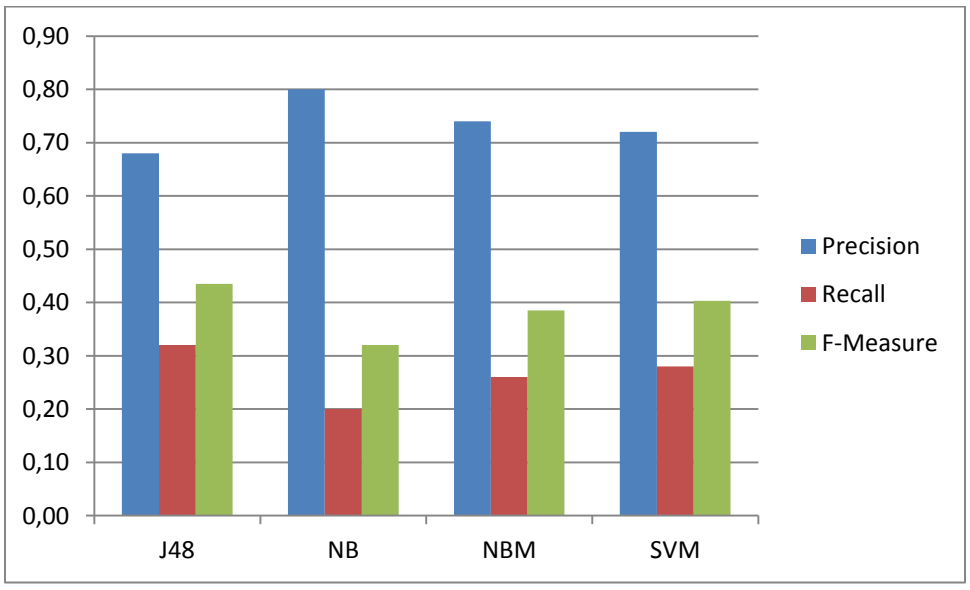


**Fig. 5.1.4. Evaluación de los clasificadores sobre el conjunto de prueba 4.**

En la Tabla 5.2, se muestran los valores de precisión que corresponden a los clasificadores utilizados con respecto a los diferentes conjuntos de prueba con los cuales se evaluó el modelo de clasificación.

**Tabla 5.2. Valores de *precisión* obtenidos por los clasificadores sobre los diferentes conjuntos de prueba.**

| *Clasificador* | *Test 1* | *Test 2* | *Test 3* | *Test 4* |
|---|---|---|---|---|
| **J48** | 0.88 | 0.78 | 0.84 | 0.68 |
| **NB** | 0.74 | 0.58 | 0.86 | 0.8 |
| **NBM** | 0.72 | 0.64 | 0.86 | 0.74 |
| **SVM** | 0.68 | 0.64 | 0.88 | 0.72 |

En la figura 5.2.1, se muestra la evaluación de precisión por cada conjunto de prueba como y su comportamiento con los diferentes clasificadores. Se puede observar que el comportamiento en cada conjunto de prueba es diferente en cada clasificador.

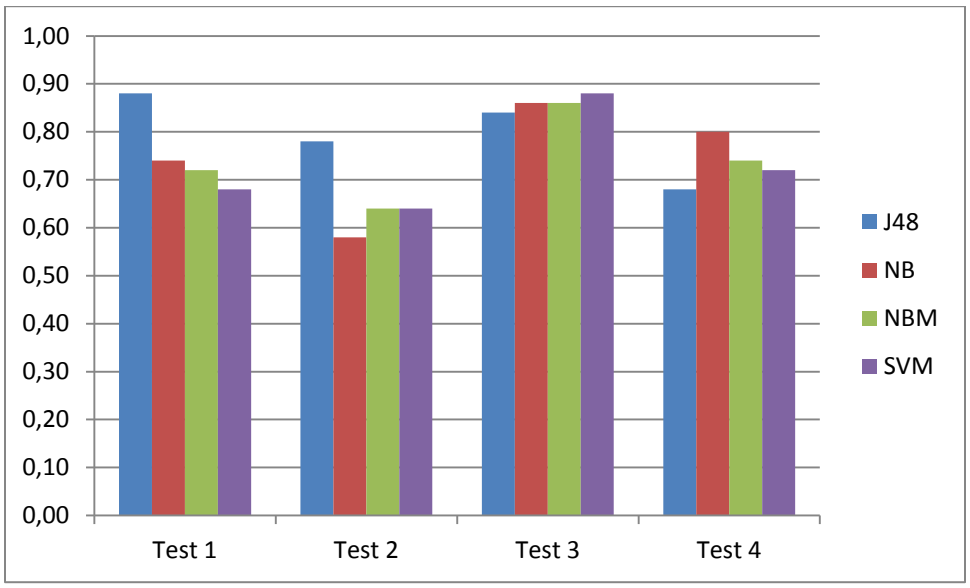


**Fig. 5.2.1. Evaluación de *precisión* sobre los diferentes conjuntos de prueba.**

En la figura 5.2.2, se muestra la evaluación de precisión por cada método de clasificación y como es el comportamiento con los diferentes conjuntos de prueba. Se puede observar que el comportamiento en cada clasificador es diferente con respecto a cada conjunto de prueba.

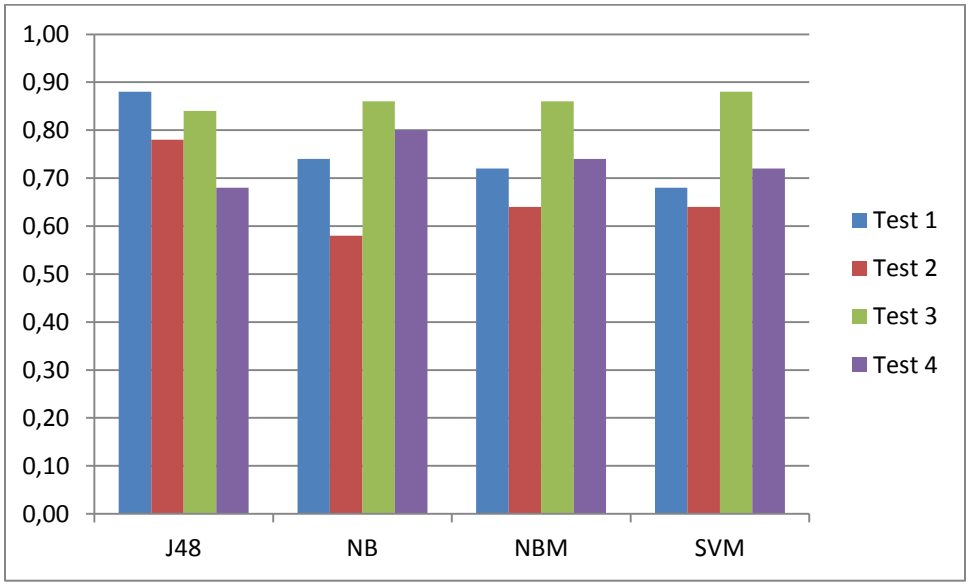


**Fig. 5.2.2. Evaluación de *precisión* sobre los diferentes métodos de clasificación.**

En la Tabla 5.3, se muestran los valores de recuerdo que corresponden a los clasificadores utilizados con respecto a los diferentes conjuntos de prueba con los cuales se evalúo el modelo de clasificación.

**Tabla 5.3. Valores de *recuerdo* obtenidos por los diferentes clasificadores sobre los diferentes conjuntos de prueba.**

| *Clasificador* | *Test 1* | *Test 2* | *Test 3* | *Test 4* |
|---|---|---|---|---|
| **J48** | 0.12 | 0.22 | 0.16 | 0.32 |
| **NB** | 0.26 | 0.42 | 0.14 | 0.2 |
| **NBM** | 0.28 | 0.36 | 0.14 | 0.26 |
| **SVM** | 0.32 | 0.36 | 0.12 | 0.28 |

En la figura 5.3.1, se muestra la evaluación del recuerdo por cada conjunto de prueba como y su comportamiento con los diferentes clasificadores. Se puede observar que el comportamiento en cada conjunto de prueba es diferente en cada clasificador. En la figura 5.3.2, se muestra la evaluación del recuerdo por cada método de clasificación y como es el comportamiento con los diferentes conjuntos de prueba. Se puede observar que el comportamiento en cada clasificador es diferente con respecto a cada conjunto de prueba.

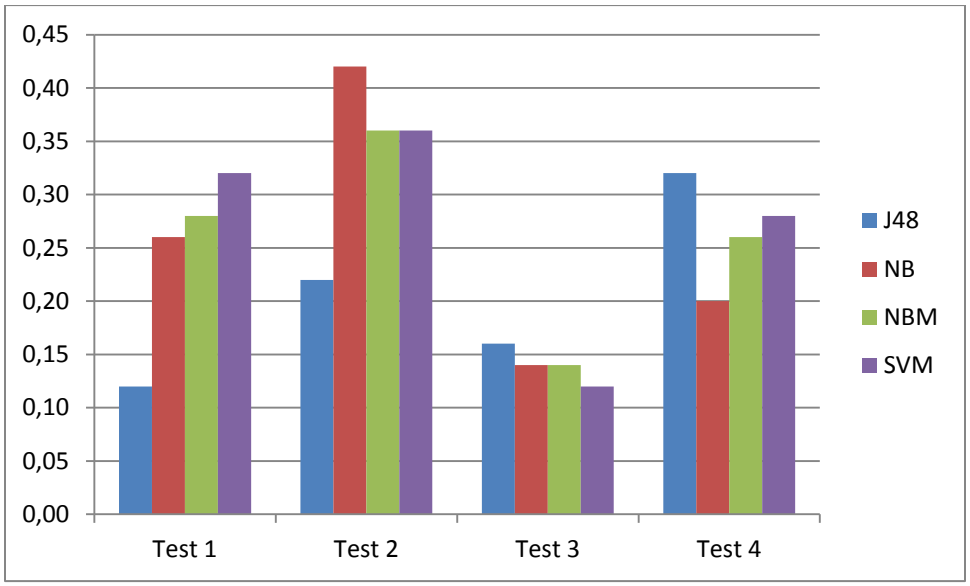


**Fig. 5.3.1. Evaluación de *recuerdo* sobre los diferentes conjuntos de prueba.**

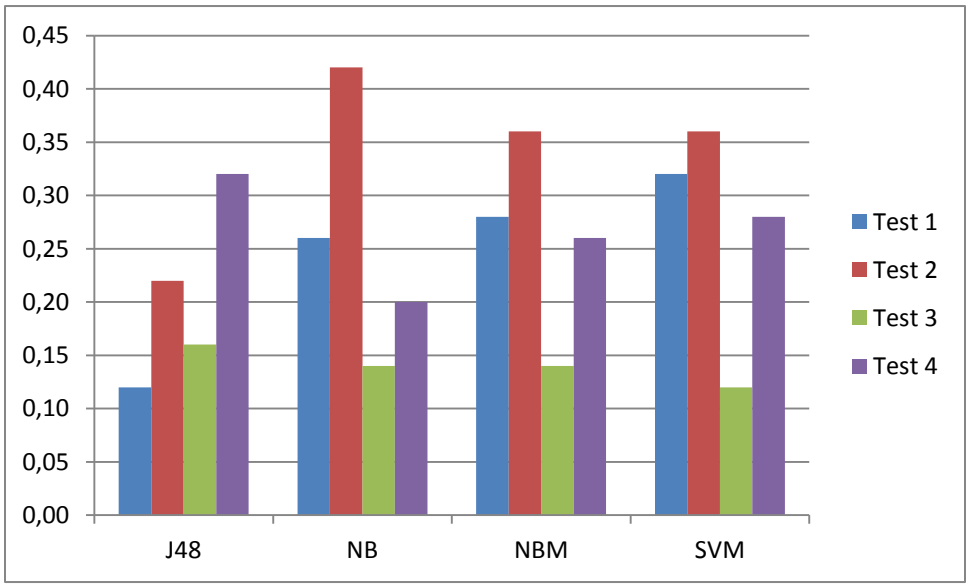


**Fig. 5.3.2. Evaluación de *recuerdo* sobre los diferentes métodos de clasificación.**

En la Tabla 5.4, se muestran los valores de F-Measure que corresponden a los clasificadores utilizados con respecto a los diferentes conjuntos de prueba con los cuales se evalúo el modelo de clasificación.

**Tabla 5.4. Valores de *F-Measure* obtenidos por los diferentes clasificadores sobre los diferentes conjuntos de prueba.**

| *Clasificador* | *Test 1* | *Test 2* | *Test 3* | *Test 4* |
|---|---|---|---|---|
| **J48** | 0.2112 | 0.3432 | 0.2688 | 0.4352 |
| **NB** | 0.3848 | 0.4872 | 0.2408 | 0.32 |
| **NBM** | 0.4032 | 0.4608 | 0.2408 | 0.3848 |
| **SVM** | 0.4352 | 0.4608 | 0.2112 | 0.4032 |

En la figura 5.4.1, se muestra la evaluación de la medida armónica F-Measure por cada conjunto de prueba como y su comportamiento con los diferentes clasificadores. Se puede observar que el comportamiento en cada conjunto de prueba es diferente en cada clasificador.

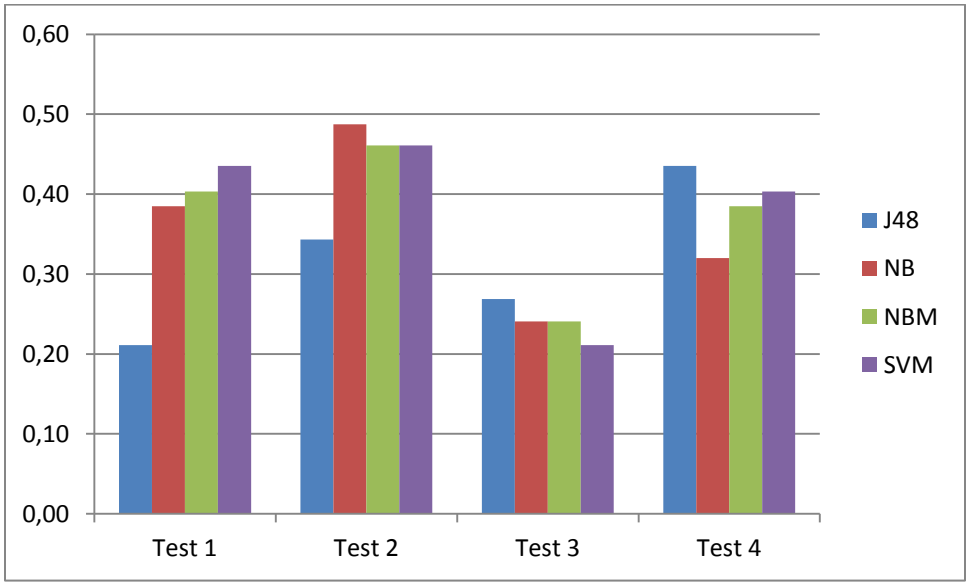


**Fig. 5.4.1. Evaluación de *F-Measure* sobre los diferentes conjuntos de prueba.**

En la figura 5.4.2, se muestra la evaluación de la medida armónica F-Measure por cada método de clasificación y como es el comportamiento con los diferentes conjuntos de prueba. Se puede observar que el comportamiento en cada clasificador es diferente con respecto a cada conjunto de prueba.

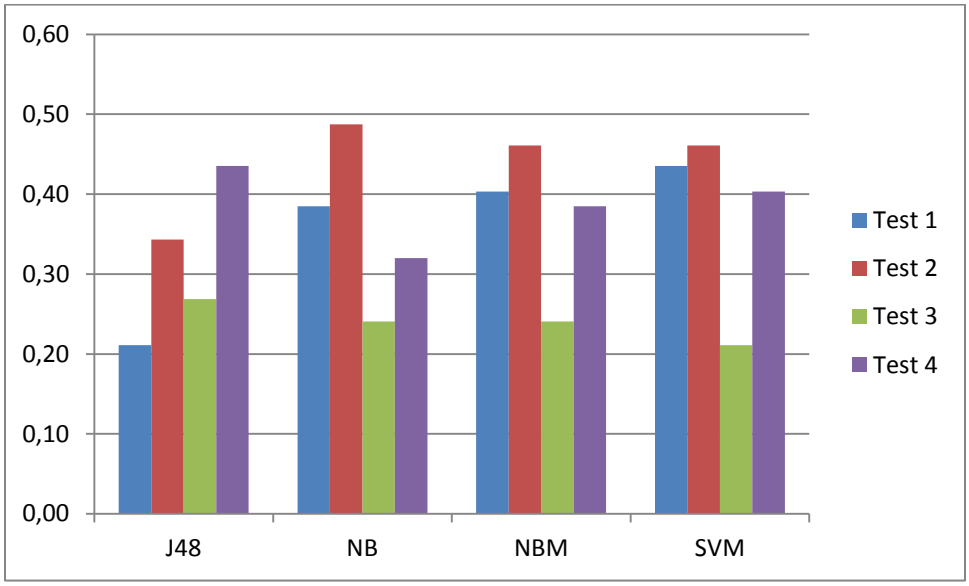


**Fig. 5.4.2. Evaluación de *F-Measure* sobre los diferentes métodos de clasificación.**

## 5.2. Resultados experimentales sobre el corpus inicial

En la Tabla 5.5, se muestran los resultados de referencia, para los diferentes tipos de clasificadores utilizados. Estos valores corresponden al porcentaje de instancias clasificadas correctamente con el corpus inicial y al enriquecer el corpus con la información descargada de la Web. En todos los casos, se determinó el contexto basándose en una bolsa de palabras que contiene la frecuencia de aparición de las palabras y se eliminaron los signos de puntuación. Para cada clasificador se utilizó un esquema de validación cruzada usando el 80% de muestras para entrenamiento (*training*) y el 20% para prueba (*test*).

**Tabla 5.5. Valores de evaluación del procedimiento de desambiguación topónimos.**

| Clasificador | Corpus Inicial | Corpus Enriquecido |
|---|---|---|
| **J48** | 68% | 71.27% |
| **NB** | 74% | 77.83% |
| **NBM** | 72% | 75.96% |
| **SVM** | 64% | 79.94% |

En la figura 5.5, se muestra la evaluación del procedimiento de desambiguación de topónimos, estos valores corresponden al porcentaje de instancias correctamente clasificadas en el corpus inicial y al enriquecer el corpus. Se puede notar que al enriquecer el corpus incrementa el porcentaje de instancias correctamente clasificadas.

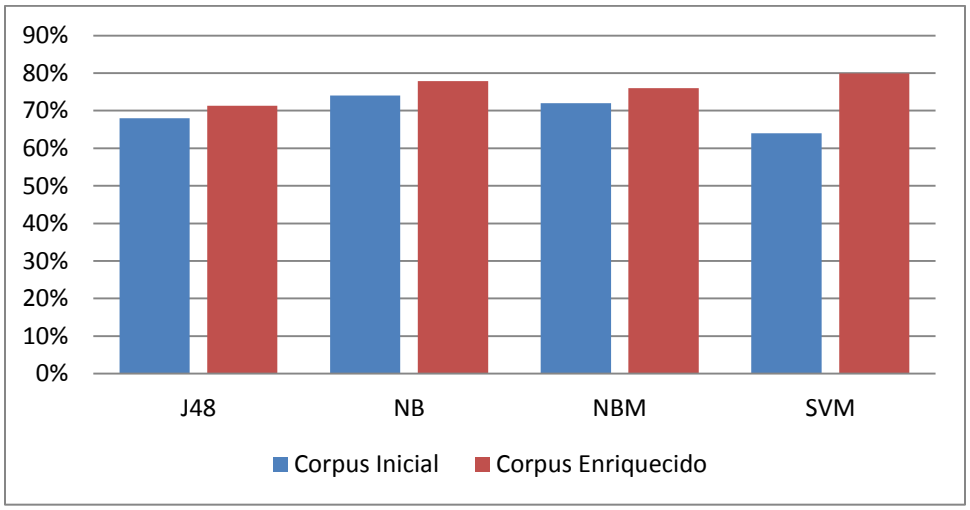


**Fig. 5.5. Evaluación del procedimiento de desambiguación topónimos.**

La Tabla 5.6, muestra los resultados de las medidas de evaluación obtenidas al realizar la desambiguación de topónimos, el valor más bajo de recuperación obtenida indica que deben incorporarse instancias negativas a los conjuntos de entrenamiento (*training*) y prueba (*testing*).

**Tabla 5.6. Valores de las medidas de evaluación del procedimiento de desambiguación topónimos.**

| Clasificador | Corpus Inicial | | | Corpus Enriquecido | | |
|---|---|---|---|---|---|---|
| | *Precisión* | *Recuerdo* | *F -Measure* | *Precisión* | *Recuerdo* | *F-Measure* |
| **J48** | 0.875 | 0.124 | 0.217 | 0.812 | 0.411 | 0.545 |
| **NB** | 0.778 | 0.221 | 0.344 | 0.824 | 0.424 | 0.563 |
| **NBM** | 0.839 | 0.161 | 0.270 | 0.893 | 0.447 | 0.595 |
| **SVM** | 0.786 | 0.213 | 0.335 | 0.833 | 0.425 | 0.562 |

La figura 5.6.1, muestra la evaluación de la precisión del procedimiento de desambiguación de topónimos, con el corpus inicial y al enriquecer el corpus. Al realizar el enriquecimiento del corpus, la precisión del clasificador aumenta. Constatando que al realizar esta etapa de la metodología de desambiguación, permite apreciar que la incorporación de información proveniente de la Web al conjunto de entrenamiento es útil.

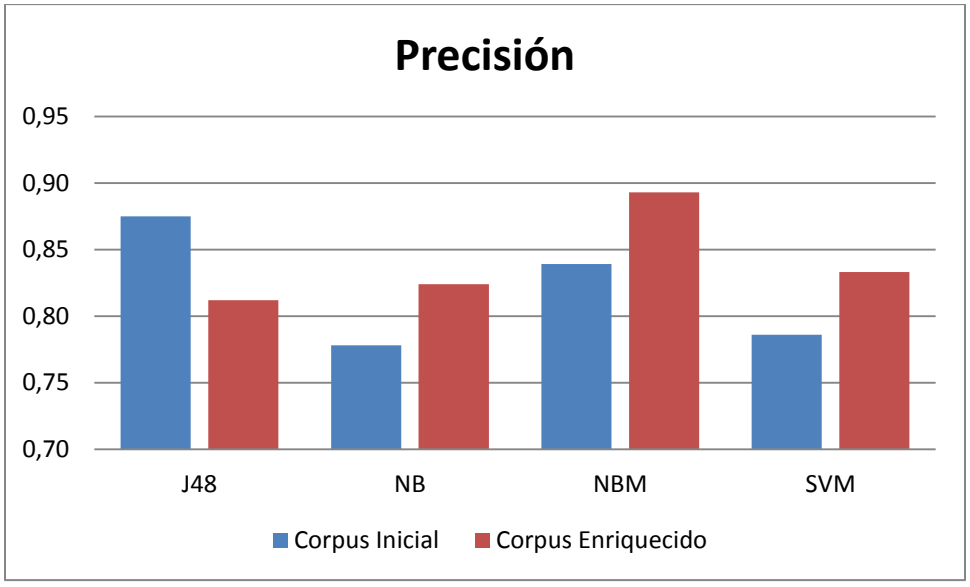


**Fig. 5.6.1. Evaluación de *precisión* del procedimiento de desambiguación topónimos.**

La figura 5.6.2, muestra la evaluación del recuerdo del procedimiento de desambiguación de topónimos, con el corpus inicial y al enriquecer el corpus. Al realizar el enriquecimiento del corpus, el recuerdo del clasificador aumenta. La figura 5.14, muestra que cuando se calcula la media armónica entre la precisión y el recuerdo, tanto para el corpus inicial como para el corpus enriquecido, se observa que incrementa. Obteniéndose así un balance entre estas dos medidas y que el hecho de enriquecer el corpus potencialmente puede ser utilizado cuando no hay suficientes instancias de entrenamiento etiquetadas para un dominio particular.

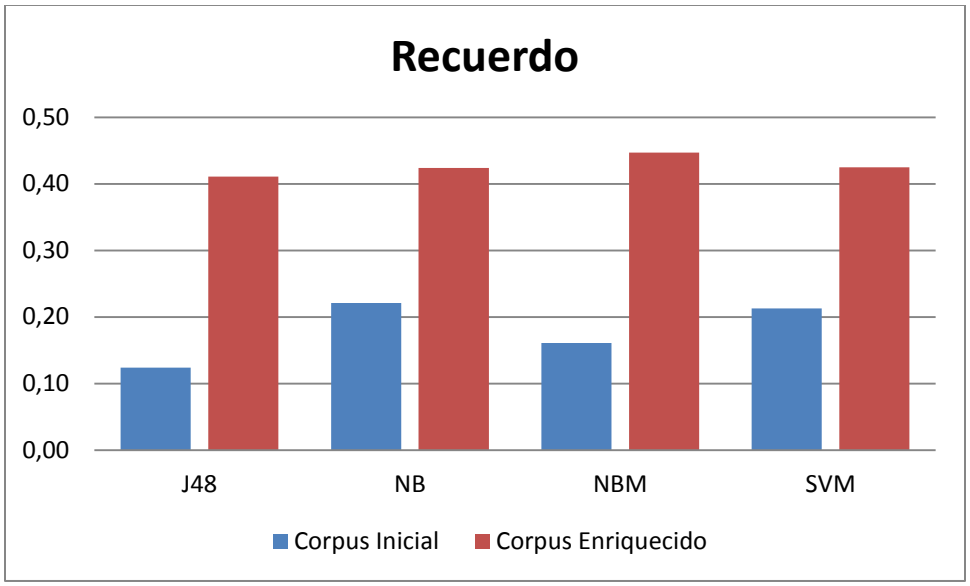


**Fig. 5.6.2. Evaluación de *recuerdo* del procedimiento de desambiguación topónimos.**

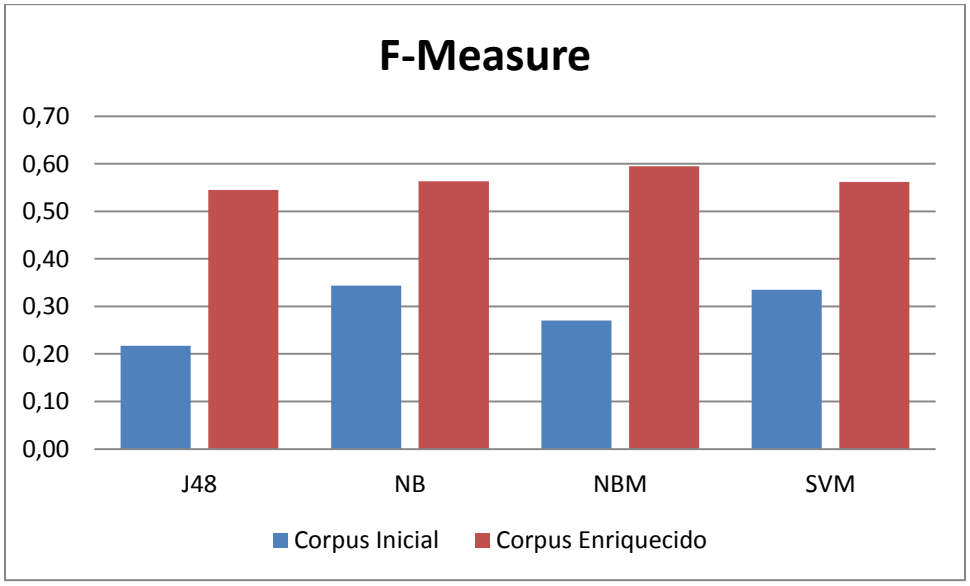


**Fig. 5.6.3. Evaluación de *F-Measure* del procedimiento de desambiguación topónimos.**

## 5.3. Discusión

Como puede observarse en la tabla 5.5, el método de clasificación que mejor se comporta es SVM. Sin embargo, es curioso que este método tenga un comportamiento pobre cuando se usa únicamente el corpus inicial para la fase de entrenamiento. Consideramos que SVM no logra capturar adecuadamente el vector de soporte dado el número reducido de muestras. En cambio el método de Naïve Bayes parece encontrar suficiente evidencia para obtener un porcentaje de clasificación aceptable. De cualquier manera, todos los clasificadores mejoran significativamente su rendimiento al enriquecer el corpus inicial.

El resultado obtenido evidencia que cuando el corpus se enriquece con la información descargada de la Web se logra incrementar la exactitud. Este hecho permite apreciar que la incorporación de información proveniente de la Web al conjunto de entrenamiento es útil para mejorar la exactitud, lo cual potencialmente puede ser utilizado cuando no hay suficientes instancias de entrenamiento etiquetadas para un dominio particular. Sin embargo, también se puede observar que existe una necesidad de aumentar el tamaño del conjunto de entrenamiento y prueba mediante la incorporación de nuevos ejemplos no etiquetados.

# CAPÍTULO 6.

# CONCLUSIONES Y TRABAJO FUTURO

Este capítulo resume las ideas, aportaciones y resultados más importantes obtenidos en este trabajo de tesis, con base en esto se describen las líneas de trabajo futuro para continuar con la investigación de esta tesis.

# Capítulo 6

# Conclusiones y trabajo futuro

La línea de investigación a la cual pertenece esta tesis es la recuperación de información geográfica, en específico se presentó una metodología para la tarea de desambiguación de topónimos, la cual consistió de: una ontología como repositorio de sentidos, un corpus enriquecido como contexto y un método desambiguador basado en un modelo de clasificación. En este capítulo final se resumen las ideas, aportaciones y resultados más importantes obtenidos en este trabajo de tesis. En la sección 6.1 se presentan las conclusiones, indicando las aportaciones principales del trabajo realizado en la sección 6.2, y en la sección 6.3 se describen las líneas de trabajo futuro que pueden realizarse a partir de la presente tesis y de la línea de investigación a la que pertenece.

## 6.1. Conclusiones

Este trabajo de tesis presenta los resultados que se obtuvieron en el campo de investigación de la recuperación de información geográfica, en específico en la tarea de desambiguación de topónimos. Este campo, que ha surgido recientemente, tiene su origen en la recuperación de información y en los sistemas de información geográfica.

Actualmente, en el área de recuperación de información geográfica aún existen muchos retos y tareas relativas a la geografía, en las cuales es necesario investigar para encontrar mejores métodos para preprocesar, buscar y manejar la información geográfica. Dichos retos se deben a que básicamente, la geografía de cada país es diferente tanto a nivel de accesibilidad, creación y generación de cartografía, así como de regulación de la misma.

Una solución para desambiguar referencias geográficas, en específico topónimos, es presentada en esta tesis. La desambiguación de topónimos constituye una de las tareas importantes dentro de la recuperación de información geográfica, debido a la alta frecuencia de aparición de topónimos en consultas GIR en la Web, y considerando además, que existen pocas investigaciones para el idioma Español realizadas sobre este tema en comparación con otros idiomas, específicamente con el idioma inglés. Por tal motivo, este trabajo se centró principalmente en la investigación y desarrollo de una metodología para la desambiguación de topónimos para el idioma Español, la cual se centra principalmente en

la construcción de una ontología como repositorio de sentidos, un corpus enriquecido como contexto y un método desambiguador basado en un modelo de clasificación.

En este trabajo se obtuvo esencialmente una ontología espacial de la República Mexicana, que sirve como repositorio de sentidos en la metodología para la tarea de desambiguación de topónimos tipo GEO/NO-GEO para el idioma Español. Un corpus supervisado que permite calcular un modelo adecuado de clasificación para esta tarea. Además de crear un corpus supervisado, se ha evaluado una metodología para la desambiguación de topónimos y se observa que los resultados fueron favorables, alrededor de un 80% de precisión para el tipo de desambiguación de topónimos realizada. Este rendimiento es principalmente debido a la alta calidad de las instancias de entrenamiento y a la incorporación de información descargada de la Web, la cual tiene garantizada una relación semántica con los topónimos a desambiguar, dados los patrones de búsqueda construidos con base a *n_gramas*.

## 6.2. Aportaciones

La metodología para la desambiguación de topónimos propuesta es la principal aportación de este trabajo de tesis. A partir de esta aportación principal se deducen las siguientes:

- Una metodología que desambigua topónimos tipo GEO/NO-GEO.
- Una ontología espacial, que incluye la representación de objetos geográficos naturales y artificiales de la República Mexicana.
- Un método de búsqueda de topónimos candidatos con base a *n_gramas.*
- Un corpus supervisado que permite calcular un modelo adecuado de clasificación para la tarea de desambiguación de topónimos.
- Un etiquetador ontológico capaz de identificar los componentes ontológicos de cualquier archivo *owl*.

## 6.3. Líneas de trabajo futuro

El enfoque presentado en este trabajo se centra en la utilización del método propuesto para la desambiguación de topónimos, sin embargo, en un futuro se podría utilizar el mismo método para la desambiguación del sentido de las palabras en general. Se propone la identificación de características, con el fin de incluir estas características en el modelo de clasificación propuesto para desambiguar topónimos e incluir una interfaz gráfica.

La ontología espacial construida fue utilizada como recurso de apoyo en la desambiguación de topónimos, sin embargo, esta ontología puede ser utilizada para desambiguar topónimos en consultas a la Web, relacionadas por ejemplo con: el ejercicio y la promoción del turismo en México, aplicaciones de cambio climático, realización de planeaciones urbanas y desarrollo de planes estratégicos de tipo económico-sociales, entre otros usos. Hasta este momento la ontología propuesta sólo comprende objetos geográficos de la República Mexicana, debido a que en otros países de habla hispana la división política varía de acuerdo a cada país, pero esta ontología puede ser extendida y llegar a tener más granularidad. Además, se plantea la incorporación de relaciones y axiomas espaciales asociados generalmente a un topónimo. La incorporación de éstos junto con la ontología realizada mejorará la ambigüedad de topónimos.

# Referencias bibliográficas


1. C. B. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M.J. van Kreveld y R. Weibel.: Spatial Information Retrieval and Geographical Ontologies an overview of the SPIRIT project. En SIGIR'02: Proc. of the 25th ACM SIGIR Conference, (2002) 387 – 388

2. C.B. Jones, A.I. Abdelmoty, y G. Fu.: Maintaining Ontologies for Geographical Information Retrieval on the Web. En ODBASE'03: Proc. of the On The Move to Meaningful Internet Systems, volume 2888 of LNCS, (2003)

3. C.B. Jones, A.I. Abdelmoty, G. Fu, y S. Vaid.: The SPIRIT Spatial Search Engine: Architecture, Ontologies and Spatial Indexing. En Proc. of the 3rd Int. Conf. on Geogr. Inform. Science, volume 3234 of LNCS, (2004) 125 – 139

4. G. Fu, C.B. Jones, y A.I. Abdelmoty.: Ontology-Based Spatial Query Expansion in Information Retrieval. En ODBASE'05: Proc. of the On the Move to Meaningful Internet Systems, volume 3761 of LNCS, (2005) 1466 – 1482

5. Buscaldi, D.: Toponym Disambiguation in Information Retrieval. PhD thesis, Universidad Politécnica de Valencia, Valencia, España (2010)

6. Pink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T.: Searching the Web: the public and their queries. Journal of the American Society for Information Science and Technology 52(3) (2001) 226–234

7. Lee, Y.K., Ng, H.T.: An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10. EMNLP '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 41–48

8. Guzmán-Cabrera, R., Rosso, P., Montes-Y-Gómez, M., Villaseñor Pineda, L., Pinto-Avendaño, D.: Semi-supervised Word Sense Disambiguation using the Web as corpus. In: Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing. CICLing '09, Berlin, Heidelberg, Springer-Verlag (2009) 256–265

9. M.F. Worboys. GIS: A Computing Perspective. CRC, 2004. ISBN: 0415283752 (2004)

10. ISO/IEC. Geographic Information – Reference Model. International Standard 19101, ISO/IEC, (2002)

11. Open GIS Consortium, Inc. OpenGIS Reference Model. OpenGIS Project Document 03-040, Open GIS Consortium, Inc., (2003)

12. Global Spatial Data Infrastructure Association. Sitio web. Fecha de consulta: Agosto de 2012. Disponible en: http://www.gsdi.org/

13. Leidner, J.L.: Toponym Resolution in Text : Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Universal Press, Boca Raton, FL, USA (2008)

14. Andogah, G.: Geographically Constrained Information Retrieval. PhD thesis, University of Groningen, Groningen,Netherlands (2010)

15. Tim Berners-Lee, James Hendler and Ora Lassila: The Semantic Web, Scientific American, (2001)

16. Kobayashi M. and Takeda, K.: Information Retrieval on the Web, ACM Computing Surveys (CSUR), Volume 32 (2000) 144-173, ISSN: 0360-0300

17. Regina M., M. Braga, Marta Mattoso, Claudia M. L. Werner: The Use of Mediation and Ontology Technologies for Software Component Information Retrieval, Proceedings of the 2001 symposium on Software reusability: putting software reuse un context, Toronto, Ontario, Canda (2001) 19-28, ISSN: 0163-5948

18. Yoel ledo Mezquita, Grigori Sidorov, A.G.: Information retrieval with word sense disambiguation for spanish. Computación y Sistemas 11 (2008) 288–300

19. Dwivedi, S.K., Rastogi, P.: Critical analysis of wsd algorithms. In: Proceedings of the International Conference on Advances in Computing, Communication and Control. ICAC3 '09, New York, NY, USA, ACM (2009) 62–67

20. Levanchkine S. & Guzmán-Arenas.: Hierarchies Measuring Qualitative Vatiables, in Gelbukh (Ed.), Computational Linguistics and Intelligent Text Processing (CICLing' 2004), Lectures Notes in Computer Science, Vol. 2945, Springer-Verlag, (2004) 262-274

21. Varelas, G., Voutsakis, E., Raftopoulou, P., Petrakis, E.G., Milios, E.E.: Semantic similarity methods in wordnet and their application to information retrieval on the web. In: Proceedings of the 7th annual ACM international workshop on Web

information and data management. WIDM ’05, New York, NY, USA, ACM (2005) 10–16

22. Andrade, L., Silva, M.J.: Relevance ranking for geographic IR. In: GIR. (2006)

23. Smith, D.A., Crane, G.: Disambiguating geographic names in a historical digital library. In: Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries. ECDL ’01, London, UK, UK, Springer-Verlag (2001) 127–136

24. Max Egenhofer, Toward the Semantic Geospatial Web, ACM-GIS 2002, A. Voisard and S.-C. Chen (eds.), (2002)

25. Jones, C.B., Purves, R., Ruas, A., Sanderson, M., Sester, M., van Kreveld, M., Weibel, R.: Spatial information retrieval and geographical ontologies an overview of the spirit project. In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR ’02, New York, NY, USA, ACM (2002) 387–388

26. R. Guha and R. McCool and E. Miller: Semantic search, WWW2003 Proceedings of the 12th international conference on World Wide Web, ACM Press, (2003) 700-709

27. Cristiano Rocha, Daniel Schwabe, Marcus Poggi Aragao: A Hybrid Approach for Searching in the Semantic Web, Proceedings of the 13th international conference on World Wide Web, (2004) 374-383, ISBN: 1-58113-844-X

28. Reiner Kraft, Farzin Maghoul, Chi Chao Chang: Y!Q: Contextual Search at the Point of Inspiration, CIKM’05, (2005)

29. Clough, P.: Extracting metadata for spatially-aware information retrieval on the internet. In: Proceedings of the 2005 workshop on Geographic information retrieval. GIR ’05, New York, NY, USA, ACM (2005) 25–30

30. Hawking, D., Craswel, N., Bailey, P., Griths k.: Measuring Search Engine Quality, Information Retrieval, Vol. 4(1), (2001) 33-59

31. Zhou Yinhua, Xie Xing, Wang Chuang, Gong Yuchang, Ma Wei-Ying,: Hybrid Index Structures for Location-based Web Search, CIKM’05, (2005)

32. Delboni, T., Borges, K., Laender, a., : In Geographical Information Retrieval, ACM-GIR’05 (2005)

33. S. Vaid., C. B. Jones, H. Joho, and M. Sanderson.: Spatio-textual Indexing for Geogrphical Search on the Web. In Proceedings of SSTD-05, the 9thSymposium on Spatial and Temporal Databases, (2005)

34. Diana Santos and Marcirio Silveira Chaves,: The place of place in geographical IR, Proceedings of SIGIR 2006, (2006)

35. Jens Graupmann and Ralf Schenkel,: GeoSphereSearch: Context-Aware Geographic Web Search, Proceedings of SIGIR 2006, (2006)

36. Arron Walker, Binh Pham, Miles Moody, Spatial Bayesian Learning Algothms for Geographic Information Retrieval, GIS'05 (2005)

37. Koo, S.O., Lim, S.Y., Lee, S.J.: Building an ontology based on hub words for information retrieval. Web Intelligence, IEEE / WIC / ACM International Conference, (2003) 466

38. Leite, M.A.A., Ricarte, I. L.: Document retrieval using fuzzy related geographic ontologies. In: Proceedings of the 2nd international workshop on Geographic information retrieval. GIR '08, New York, NY, USA, ACM (2008) 47–54

39. Naveiras, D.S.: Técnicas de indexación y recuperación de documentos utilizando referencias geográficas y textuales. PhD thesis, Universidad de Coruña, Coruña, España (2009)

40. Lopez, A., Somodevilla, M.J., Vilarino, D., Pineda, I.H., De Celis, C.: Toponym disambiguation by ontology in spanish: Geographical proximity between place names in the same context. In: AISS: Advances in Information Sciences and Service Sciences. (2012) 282–289

41. David A. Smith and Gideon S. Mann. Bootstrapping toponym classifiers. In HLT-NAACL 2003 workshop on Analysis of geographic references, Morristown, NJ, USA, (2003) 45-49. Association for Computational Linguistics. doi: http://dx.doi.org/10.3115/1119394.1119401

42. Eneko Agirre and German Rigau. Word sense disambiguation using conceptual density. In 16th Conference on Computational Linguistics (COLING '96), Copenhaghen, Denmark, (1996) 16-22

43. Paolo Rosso, Francesco Masulli, Davide Buscaldi, Ferran Pla, and Antonio Molina. Automatic noun sense disambiguation. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, 4th International

Conference, volume 2588 of Lecture Notes in Computer Science, Springer, Berlin, (2003) 273-276

44. Navigli, R.: Word sense disambiguation: a survey. ACM COMPUTING SURVEYS 41(2) (2009) 1–69

45. Mark Sanderson. Word sense disambiguation and information retrieval. In SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, 1994. Springer-Verlag New York, Inc (1994) 142-151

46. Mark Sanderson. Retrieving with good sense. Information Retrieval, 2(1), (2000) 49-69

47. Christopher Stokoe, Michael P. Oakes, and John Tait. Word Sense Disambiguation in Information Retrieval revisited. In SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, New York, NY, USA, ACM. doi: 10.1145/860435.860466, (2003) 159-166

48. Nicola Stokes, Yi Li, Alistair Mo ffat, and Jiawen Rong. An empirical study of the effects of nlp components on geographic ir performance. International Journal of Geographical Information Science, 22(3), (2008) 247-264

49. Bucher, B., P. Clough, H. Joho, R. Purves, y A. K. Syed. 2005. Geographic IR Systems: Requirements and Evaluation. En Proceedings of the 22nd International Cartographic Conference (2005)

50. García Cumbreras, M.A., L.A. Ureña-López, F. Martínez Santiago, y J.M. Perea Ortega. 2007. BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006. LNCS of Springer-Verlag (2006)

51. Rada R., Mill H., Bicknell E. and Blettner M. : Development an Application of a Metric on Semantic Nets, in IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, (1989 ) 17-30

52. Sussna M. : Word Sense Disambiguation for Free-text Indexing Using a Massive Semantic Network, in Proceedings of the Second International Conference on Information and knowledge Management. Arlington, Virginia.

53. Resnik P. : Disambiguating Noun Groupings with Respect to WordNet Senses, in Proceedings of the Third Workshop on Very Large Corpora, MIT (1995)

54. Mark Sanderson, Kalervo Järvelin, James Allan, Peter Bruza (Eds.): SIGIR 2004: Proceedings of the 27th Annual International ACM SIGIR Conference on Rerearch Development in Information Retrieval, Sheffield (2004) 25-29, ISBN 1-58113-881-4

55. Ray R. Larson,: Geographic Information Retrieval and Spatial Browsing, University of California, Berkeley. (2007)

56. Ray R. Larson, Patricia Frontiera, Geographic Information Retrieval (GIR) Ranking Methods for Digital Libraries, Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Labraries, Tuscon, AZ, USA (2004) 415-415

57. Ricardo Baeza-Yates and Berthier Ribeiro-Neto.: Modern Information Retrieval. ACM Press, New York, NY, (1999)

58. Ian H. Witten, Timothy C. Bell, and Craig G. Neville. Indexing and Compressing Full-Text Databases for CD-ROM. J. Information Science, (1992) 17:265-271

59. Gerard Salton and Michael Lesk.: Computer evaluation of indexing and text processing. J. ACM, (1968) 15(1):8-36

60. Ferrés, D. y H. Rodríguez. 2007. TALP at GeoCLEF 2007: Using Terrier with Geographical Knowledge Filtering. En Working Notes of the Cross Language Evaluation Forum (2007)

61. Larson, R.R. 2007. Cheshire at GeoCLEF 2007: Retesting Text Retrieval Baselines. En Working Notes of the Cross Language Evaluation Forum (2007)

62. García Cumbreras, M.A., L.A. Ureña-López, F. Martínez Santiago, y J.M. Perea Ortega. 2007. BRUJA System. The University of Jaén at the Spanish task of QA@CLEF 2006. LNCS of Springer-Verlag (2006)

63. Dan Moldovan, Marius Pasca, Sanda Harabagiu, and Mihai Surdeanu.: Performance issues and error analysis in an open-domain question answering system. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, New York, USA, (2003)

64. Gruber T.R.: "Towards Principles for the Design of Ontologies Used for Knowledge Sharing". In Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer, The Netherlands, (1993)

65. Gruber T.R.: "A Translation Approach to Portable Ontology Specifications". Knowledge Acquisition, 5(2), (1993) 199 – 220

66. Haarslev V., Lutz C., Moller R.: “Foundations of Spatioterminological Reasoning with Description Logics”, Proceedings of the sixth Int. Conf. On Principles of Knowledge Representation and Reasoning (KR'98), A.G. Cohn et al, (1998) 112-123

67. Spaccapietra S., Cullot N., Parent C., Vangenot C.: “On Spatial Ontologies”, 6th Brazilalian Symposium on GeoInformatics, GeoInfo, Campos do Jordao, Brazil, Noviembre 22-24, (2004)

68. Tanasescu V., Gugliotta A., Domingue J., Davies R., Gutiérrez-Villarías L., Rowlatt M. Richardson M., Stincic S.: “A Semantic Web Servicies GIS based Emergency Management Application”, International Semantic Web Conference, Athens, GA, USA, (2006) 959-966

69. Mooney, R. J.: Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. In Proceedings of the 1996 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1996 82–91

70. Boser B.E., Guyon I.M., and Vapnik V.N: A training algorithm for optimal margin classifiers. In Proceedings of the 5th Annual Workshop on Computational Learning Theory (Pittsburgh, PA), 1992 144–152

# Anexo A
# Conjunto de etiquetas de TreeTagger

En este anexo, se muestra la abreviatura de etiquetas de la parte de la oración (etiqueta POS) que da como resultado el etiquetador TreeTagger y su correspondiente parte de la oración de cada etiqueta. Cada etiqueta POS corresponde al archivo de parámetros del idioma español con codificación en UTF-8. La tabla A1, describe este anexo.

**Tabla A1. Conjunto de etiquetas TreeTagger.**

| Etiqueta POS | Correspondiente a: |
|---|---|
| ACRNM | Siglas (ISO,CEI) |
| ADJ | Adjetivos (mayores, mayor) |
| ADV | Adverbio (muy, demasiado, cómo) |
| ALFP | Letras en plural del alfabeto (as, aes, bes) |
| ALFS | Letras en singular del alfabeto (a, b, c) |
| ART | Artículo (un, las, la, unas) |
| BACKSLASH | Barra diagonal inversa (\) |
| CARD | Cardinales |
| CC | Conjunciones coordinadas o coordinantes (y, o) |
| CCAD | Conjunción coordinada adversativa (pero) |
| CCNEG | Conjunción coordenada negativa |
| CM | Coma (,) |
| CODE | Código alfanumérico |
| COLON | Dos puntos (:) |
| CQUE | Que (como conjunción) |
| CSUBF | Conjunción subordinada que introduce cláusulas (apenas) |
| CSUBI | Conjunción subordinada que introduce cláusulas infinitas (al) |

| | |
|---|---|
| CSUBX | Conjunción subordinada bajo especificado para el tipo subordinado (aunque) |
| DASH | Guión (-) |
| DM | Pronombres demostrativos (ésas, ése, esta) |
| DOTS | Etiqueta POS para "..." |
| FO | Formula |
| FS | Signo de puntuación, punto final. |
| INT | Pronombres interrogativos (quiénes, cuántas, cuanto) |
| ITJN | Exclamación (oh, ja) |
| LP | Paréntesis izquierdo ("(", "[") |
| NC | Sustantivos comunes (mesas, mesa, libro, ordenador) |
| NEG | Negación |
| NMEA | Sustantivos métricos |
| NMON | Nombre de meses |
| NP | Sustantivos propios |
| ORD | Ordinales (primer, primeras, primera) |
| PAL | Acrónimo de una palabra formada por “a” y “el” |
| PDEL | Acrónimo de una palabra formada por “de” y “el” |
| PE | Palabra extranjera |
| PERCT | Signo de porcentaje (%) |
| PNC | Palabra no clasificada |
| PPC | Pronombres personales clíticos (le, les) |
| PPO | Pronombres posesivos (mi, su, sus) |
| PPX | Clíticos y pronombres personales (nos, me, nosotras, te, sí) |
| PREP | Preposición negativa (sin) |
| PREP | Preposición |
| PREP/DEL | Preposición compleja "después del" |
| QT | Símbolo de citación (" ' `) |
| QU | Cuantificadores (sendas, cada) |

| | |
|---|---|
| REL | Pronombre relativos (cuyas, cuyo) |
| RP | Paréntesis derechos (")", "]") |
| SE | Se (como partícula) |
| SEMICOLON | Punto y como (;) |
| SLASH | Diagonal (/) |
| SYM | Símbolos |
| UMMX | Unidades de medida (MHz, km, mA) |
| VCLIger | Clíticos del gerundio de un verbo |
| VCLIinf | Clíticos de un verbo en infinitivo |
| VCLIfin | Clíticos de un verbo en finito |
| VEadj | Verbo estar. Pasado participio |
| VEfin | Verbo estar. Finito |
| VEger | Verbo estar. Gerundio |
| VEinf | Verbo estar. Infinitivo |
| VHadj | Verbo haber. Pasado participio |
| VHfin | Verbo haber. Finito |
| VHger | Verbo haber. Gerundio |
| VHinf | Verbo haber. Infinitivo |
| VLadj | Verbo léxico. Pasado participio |
| VLfin | Verbo léxico. Finito |
| VLger | Verbo léxico. Gerundio |
| VLinf | Verbo léxico. Infinitivo |
| VMadj | Verbo modal. Pasado participio |
| VMfin | M Verbo modal. Finito |
| VMger | Verbo modal. Gerundio |
| VMinf | Verbo modal. Infinitivo |

| | |
|---|---|
| VSadj | Verbo ser. Pasado participio |
| VSfin | Verbo ser. Finito |
| VSger | Verbo ser. Gerundio |
| VSinf | Verbo ser. Infinitivo |