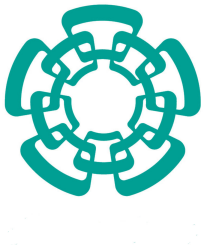

Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional

Unidad Tamaulipas

# Método para la desambiguación de topónimos con base en el contexto de textos en Español

Tesis que presenta:

**Shanel Daniela Reyes Palacios**

Para obtener el grado de:

**Maestro en Ciencias en Ingeniería y Tecnologías Computacionales**

Dr. Edwyn Javier Aldana Bobadilla, Co-director
Dr. Iván López Arévalo, Co-director

Cd. Victoria, Tamaulipas, México. Septiembre, 2020

La tesis presentada por Shanel Daniela Reyes Palacios fue aprobada por:

---

Dr. Hiram Galeana Zapién

Dr. Alejandro Molina Villegas

Dr. Iván López Arévalo, Director

Dr. Edwyn Javier Aldana Bobadilla, Director

Cd. Victoria, Tamaulipas, México., 30 de Septiembre de 2020

A mis padres

# Agradecimientos


- Gracias a Dios por la sabiduría y fortaleza que me dio para no abandonar cuando todo se veía difícil.
- A mis padres porque sin ellos no sería la persona que soy, porque a pesar de la distancia siempre están cerca para guiarme, aconsejarme, apoyarme, confiar en mí y mis decisiones. Pero sobre todo por amarme.
- A mi hermano por ser mi confidente, mi consejero y mi mejor amigo, por estar conmigo en cada etapa.
- A mi novio por estar conmigo en cada momento, por todo el amor que me ha dado, la confianza, los consejos y la paciencia. Por ser un ejemplo de seguridad y fortaleza, por salir adelante a pesar de las dificultades y por motivarme a hacer lo mismo. Gracias por continuar a mi lado a pesar de todo, eres un ejemplo para mí.
- A mi hermana Carolina por confiar en mí en todo momento, por apoyarme a pesar de no estar de acuerdo con todas mis decisiones. A Kile por creer en mí y mi capacidad, y por hacerme sentir que puedo con todo.
- Al Dr. Edwyn por compartir su conocimiento conmigo, por ayudarme a creer en mí y darme la seguridad y confianza que me falta. Porque gracias a él fue posible terminar este proyecto a tiempo, le agradezco por su tiempo y por cada observación.
- Al Dr. Iván por guiarme, por brindarme su confianza y paciencia, por dedicar parte de su tiempo para enseñarme y explicarme las cosas.
- Al Dr. Alejandro Molina por su apoyo y dirección durante mi estancia en el CentroGeo en Mérida.
- A Mogui, por apoyarme, por aconsejarme, por estar para mí, por compartir tus ideas y pensamientos, tus alegrías y tus tristezas, por convencerme de comenzar esta etapa y por terminarla juntos. Tú te has convertido en mi mejor amigo.
- A mis compañeros de los que aprendí a no rendirme, en especial a Barrón y Fer, por la confianza, las experiencias vividas, las risas y lágrimas compartidas. Por hacerme sentir parte de algo especial.
- Agradezco al CINVESTAV por darme la oportunidad de estudiar un posgrado y al CONACYT por el apoyo económico durante estos dos años.

# Índice General

# Índice de Figuras

# Índice de Tablas

# Índice de Algoritmos

# Publicaciones

Edwin Aldana-Bobadilla, Alejandro Molina-Villegas, Ivan Lopez-Arevalo, Shanel Reyes-Palacios, Victor Muñiz-Sanchez, Jean Arreola. Adaptive Geoparsing Method for Toponym Recognition and Resolution in Unstructured Text. Remote Sensing. Editorial MDPI. 2020. En revisión.

Reyes-Palacios, S., Aldana-Bobadilla, E., Lopez-Arevalo, I., & Molina-Villegas, A. (2019, November). Georeference Assignment of Locations based on Context. 1st International Conference on Geospatial Information Sciences (Vol. 13, pp. 39-46).

Shanel Daniela Reyes Palacios, Edwin Aldana Bobadilla, Ivan Lopez-Arevalo. Asignación de georreferencias a entidades nombradas de localidad con base en el contexto del texto. Avances en Ciencias en Ingeniería y Tecnologías Computacionales - TopTamaulipas 2019 - Encuentro Estatal de Estudiantes Destacados en Tecnologías de Información, Cd. Victoria, Tamps., México, 25-27 de septiembre, 2019. pags 60-62. ISBN 978-607-9023-62-1.

# Resumen

## Método para la desambiguación de topónimos con base en el contexto de textos en Español

por

**Shanel Daniela Reyes Palacios**
Unidad Tamaulipas
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2020
Dr. Edwyn Javier Aldana Bobadilla, Co-Director
Dr. Iván López Arévalo, Co-Director

La Recuperación de Información es un área enfocada en recuperar documentos relacionados con las consultas de un usuario, estos documentos comúnmente contienen nombres de localidades. Esta información la solicitan los usuarios con diversos propósitos, uno de ellos es resolver una necesidad de información sobre cierta región o localidad. La extracción de elementos asociados a localidades es posible a través de Geoparsing. Geoparsing se compone de dos tareas principales: reconocimiento de localidades y resolución de localidades. Dentro de la resolución de localidades se deriva otro problema, la ambigüedad en el nombre de localidades.

En este trabajo de investigación se propone un método de desambiguación a través del cual es posible conocer dónde está ubicada cada localidad mencionada en un texto en Español. Este método está basado en la manera en la que los humanos resolvemos el problema de ambigüedad, mediante elementos del contexto del texto. Para conseguirlo se hizo uso de un Reconocedor de Entidades Geográficas especializado en México, un gazetteer y un motor de inferencia.

La evaluación realizada demuestra que el método de desambiguación asigna la mayoría de las coordenadas geográficas a cada localidad en un rango de 0 a 5 kilómetros de las coordenadas geográficas reales. Además presenta un 90.8% de precisión en la asignación de categorías geográficas a las localidades.

# Abstract

## A context-based method for the disambiguation of toponyms from text in Spanish

by

**Shanel Daniela Reyes Palacios**
Cinvestav Tamaulipas
Research Center for Advanced Study from the National Polytechnic Institute, 2020
Dr. Edwyn Javier Aldana Bobadilla, Co-advisor
Dr. Iván López Arévalo, Co-advisor

Information Retrieval is an area from computing focused on retrieving documents related to a user's queries, these documents commonly contain names of localities. This information is requested by users for several purposes, one of them is solve an information need from a region or locality. The extraction of elements associated with locations is possible through GeoParsing. Geoparsing is constituted of two main tasks: location recognition, and location resolution. Within location resolution other problem arises, the ambiguity in the name of locations.

This research work proposes a disambiguation method through which it is possible to know where each locality mentioned in a text in Spanish is located. This method is based on the manner how humans solve the problem of ambiguity using elements of the context of text. To achieve this, a specialized Geographical Named Entity Recognizer for Mexico, a gazetteer and an inference engine were used.

The evaluation of this proposal shows that the disambiguation method assigns most of the geographic coordinates to each location in a range of 0 to 5 kilometers from the real geographic coordinates. Also, it has a 90.8% accuracy in assigning geographic categories to localities.

# 1

# Introducción

En la actualidad los sistemas de cómputo han permitido generar grandes volúmenes de información como resultado de las actividades cotidianas de las personas y las organizaciones en diferentes ámbitos, que van desde la industria, el entretenimiento, la ciencia, entre muchas otras. Típicamente esta información es clasificada en dos grandes grupos: estructurada y no estructurada. La primera es el resultado de procesos cuyos datos satisfacen una estructura predefinida, como por ejemplo una hoja de cálculo o un base de datos relacional [14]. Mientras que la segunda no satisface una estructura en particular y es el producto de la evolución de las tecnologías de información, que han permitido generar datos en diferentes formatos, como imágenes, audio, videos, texto. Esto como consecuencia de la interacción de personas y organizaciones en sitios web, redes sociales, blogs entre muchos otros [15].

El trabajo desarrollado en esta tesis está enfocado a información textual, la cual ha dado origen a áreas especializadas como la minería de texto. Ésta es un área de la minería de datos enfocada al procesamiento y análisis de recursos textuales provenientes de diversas fuentes como sitios web,

correos electrónicos, comentarios en redes sociales, sitios de noticias, entre muchos otros [15]. Permite explorar grandes cantidades de datos para buscar patrones, encontrar relaciones y extraer conocimiento [35, 36].

Diversos autores (Abelleira *et al.* [1], Montes et al. [36] y Contreras [7]) dividen el proceso de minería de texto en cinco etapas principales:

1. Etapa de recolección: Se realiza la recopilación de datos de diferentes recursos, tales como sitios web, correos electrónicos, entre otros.

2. Etapa de preprocesamiento: En esta etapa el texto es transformado a una representación manejable por los algoritmos de la computadora [36].

3. Etapa de limpieza de datos: Aquí se elimina la información innecesaria o no deseada del texto.

4. Etapa de tokenización: Aquí se divide el texto en entidades significativas (palabras, oraciones, etc.) considerando los espacios en blanco presentes y las puntuaciones.

5. Etapa de extracción de características: En esta etapa los datos preprocesados previamente son analizados con el objetivo de encontrar en ellos patrones de interés o nuevo conocimiento [36].

Existen diversas tareas básicas de la minería de texto, entre las cuales destacan: agrupación, generación de resúmenes, categorización y las relacionadas con Recuperación de Información (IR por sus siglas del Inglés) y Extracción de Información (IE por sus siglas del Inglés). Los métodos de IR se enfocan en proporcionar al usuario información contenida en documentos, solicitadas a partir de consultas. La IE transforma información contenida en una colección de textos de manera que su análisis sea más simple. Este proceso consiste en separar secciones de texto de interés, extraer información relevante de cada una de las secciones y finalmente reunir esa información [12]. La IE satisface las necesidades del usuario sobre extracción de información desde texto, las cuales son cada vez más especializadas.

Muchas de las tareas del proceso de Extracción de Información se apoyan en técnicas de Procesamiento de Lenguaje Natural (NLP, por sus siglas del Inglés), las cuales permiten analizar y representar textos con el propósito de emular algunas tareas propias de los seres humanos respecto al lenguaje, tales como comprensión de texto escrito, reconocimiento y comprensión de voz, traducción, resumen de texto, entre otras, para una variedad de tareas o aplicaciones [41].

Uno de los objetivos de los métodos de IR es la obtención de información asociada a personas, objetos, lugares y sus posibles relaciones [14]. Para lograr esto hace uso del Reconocimiento de Entidades Nombradas (NER, por sus siglas del Inglés). NER es una tarea clave para la extracción de información que consiste en identificar de forma automática entidades en colecciones de texto no estructurado. Entre las entidades nombradas que podemos encontrar mediante un sistema NER se encuentran las entidades de localidad[1] [22], éstas son las que permitirán hacer la asociación entre el texto y el topónimo que se está mencionando en él. En la actualidad mucha de la información generada y buscada está fuertemente relacionada con lugares geográficos (topónimos). Existe la necesidad de tomar en cuenta a los topónimos cuando se genera información para contextualizar al usuario del texto.

En el estado del arte la identificación de topónimos mediante NER es de "grano grueso", es decir, no se identifican topónimos muy específicos o de ámbito local pequeño, sólo se identifican los topónimos más importantes. Más aún, no se distingue si una misma cadena de caracteres hace referencia a diferentes topónimos, lo que provoca ambigüedad en el nombre de topónimos. Por ejemplo "en el municipio de Guerrero vive doña Julia, originaria de Acapulco, Guerrero", donde se identifica a Guerrero como un mismo topónimo.

Muchas veces los topónimos sólo son representados e identificados por una etiqueta alfanumérica. Para facilitar las tareas de análisis de texto, es necesario y deseable que se asocien coordenadas geográficas a topónimos. La tarea de relacionar topónimos con zonas geográficas se denomina georreferenciación, la cual consiste en asignar coordenadas geográficas a los elementos naturales

[1] En el resto del documento se empleará la palabra topónimo para hacer referencia a localidad y lugar.

o artificiales que conforman el territorio. Es decir, se asignan coordenadas geográficas a los puntos necesarios para definir el objeto a georreferenciar [31]. La georreferenciación ha cobrado relevancia porque permite conocer la ubicación de cualquier porción de la superficie terrestre y de cualquier objeto sobre ella, y constituye una herramienta fundamental para la realización de Sistemas de Información Territorial [9]. La georreferenciación es una tarea costosa en tiempo, esfuerzo y dinero, ya que su realización requiere de personal especializado para llevarla a cabo, empleando distintos recursos y herramientas (ontologías, gazetteers, entre otras) [35].

Últimamente uno de los recursos más empleados en el reconocimiento y desambiguación de topónimos son los gazetteers digitales, los cuales son recursos que facilitan la georreferenciación resolviendo, de manera muy general, la desambiguación [25]. Un gazetteer digital es un diccionario de nombres geográficos que cuenta con cuatro componentes principales: 1) nombre del topónimo incluyendo variantes del mismo, 2) ubicación, incluyendo coordenadas geográficas que representan un punto, 3) tipo, una categoría que se le asigna al topónimo, y 4) nivel jerárquico, es decir, información del topónimo para determinar su lugar en la jerarquía geopolítica, en el caso de México, por ejemplo, los estados tienen municipios y estos a su vez localidades [25].

En este trabajo de investigación se propone un método para desambiguar topónimos de grano fino mencionados en un texto en Español. Esto enriquece el texto adicionando, sin ambigüedad, una capa de georreferenciación a cada topónimo. Para ello es necesario identificar entidades que hacen alusión a topónimos y asignarles una georreferencia (latitud y longitud). Esto representa un desafío que va más allá de buscar las entidades a referenciar en un gazetteer. Dadas las múltiples combinaciones de estructuras de lenguaje natural, existe una alta probabilidad de caer en ambigüedades cuando la asignación de georreferencia se realiza de forma automática. Es por esto que este trabajo se enfoca en minimizar posibles ambigüedades en cuanto a los topónimos mencionados en un texto.

## 1.1 Planteamiento del problema

Un problema interesante en Recuperación de Información es aquel en el cual se requiere extraer aquellos elementos del texto que hacen alusión a topónimos con el propósito de inferir sus propiedades geográficas (típicamente en términos de latitud y longitud). Este problema es conocido como *Geoparsing*. Un primer paso en este proceso es la identificación de topónimos usando, comúnmente, técnicas NER enfocadas exclusivamente al reconocimiento de topónimos potenciales; dichas técnicas son denominadas GNER por sus siglas del Inglés (Geographic Named Entity Recognition).

Para contextualizar el problema, considérese el siguiente texto:

> A algunos de estos sitios se les llama coloquialmente San Pancho, como sucede con San Francisco, un pueblo costero en Nayarit o con San Francisco la capital mundial del sombrero en Guanajuato.

*San Francisco (Nayarit)* y *San Francisco (Guanajuato)* serán los topónimos reconocidos a través de GNER, los cuales serán utilizados para el paso posterior que consiste en inferir sus propiedades geográficas. Este paso representa un desafío importante, ya que en este punto suelen presentarse una gran variedad de ambigüedades respecto a las propiedades geográficas de los topónimos reconocidos. En el caso del texto antes mencionado, los topónimos *San Francisco (Nayarit)* y *San Francisco (Guanajuato)* son homónimos pero hacen alusión a topónimos geográficamente distintos, este hecho es denominado como *ambigüedad geográfica*.

Existen varios enfoques que intentan resolver este inconveniente, recurriendo al uso de diccionarios geográficos (gazetteer) (ver Sección 2.7). Mediante un gazetteer es posible resolver los casos de ambigüedad. Sin embargo, dada la naturaleza finita del gazetteer, no siempre es posible encontrar todos los topónimos involucrados en la ambigüedad. Por lo tanto, aquellas ambigüedades que incluyen topónimos muy específicos o de grano fino (*Restaurante el Danubio, Parque las Flores, Calle Amazonas, etc.*) son prácticamente imposibles de resolver usando solamente gazetteers.

Desde la perspectiva de los humanos las ambigüedades como la anterior se resuelven a través de información contextual implícita en el texto. Dicha información permite deducir reglas para inferir las propiedades geográficas más probables de los topónimos involucrados. Esto plantea las siguientes preguntas de investigación:

- *¿Cómo se podría emular el proceso humano de desambiguación con base en el contexto de un texto para asignar propiedades geográficas a un conjunto de topónimos?*
- *¿Será este proceso aplicable para resolver ambigüedades que involucren topónimos de grano fino?*

A partir de estas preguntas, en la siguiente sección se presenta la hipótesis de este trabajo de investigación.

## 1.2 Hipótesis

*Es posible asignar propiedades geográficas a topónimos presentes en un texto en Español mediante un método de asociación de grano fino, tomando en cuenta el contexto del texto.*

## 1.3 Objetivos

A continuación se presentan los objetivos de este trabajo de tesis que permitirán confirmar la hipótesis anterior.

### General

Obtener un método de desambiguación que permita asignar propiedades geográficas de grano fino a topónimos presentes en un texto en Español.

### Particulares

- A partir de un texto en Español, diseñar una estrategia para identificar las capas del contexto de dicho texto.
- Definir un mecanismo para la asociación de topónimos en las diferentes capas de un contexto.
- Desarrollar una estrategia de asignación de latitud y longitud a topónimos ambiguos.

## 1.4 Metodología de desarrollo

En esta sección se describe de manera general cada una de las actividades necesarias para el desarrollo de este trabajo de investigación. La metodología está determinada por cuatro etapas ilustradas en la Figura 1.1. A continuación se describen estas etapas.

- **Definición del problema**

  Esta primera etapa está compuesta por dos actividades; la primera es la identificación del problema, para ello se plantea conocer todo lo relacionado con el tema. El objetivo principal es analizar el problema, identificar las soluciones actuales (en caso de existir) y conocer la importancia de darle solución. La segunda actividad es el estudio del estado del arte, esto permitió ubicar el avance que se tiene hasta el momento en relación con el tema que se pretende resolver.

- **Definición del método**

  Una vez identificado el tema y la forma de abordarlo, fue necesario definir el método de solución, para ello se tomaron en cuenta algunos elementos:

  - Características del corpus de entrada
  - Tareas necesarias para eliminar inconsistencias en el texto
  - Si el texto de entrada está etiquetado

- La manera de cómo se pueden identificar las partes de un contexto
- La manera de cómo asociar los topónimos de una capa con otra
- La manera de cómo implementar la identificación del contexto
- La manera de cómo representar la evidencia de un texto
- La manera de cómo seleccionar la latitud y longitud a un topónimo con base en sus propiedades.

- **Implementación del método**

  A partir de la etapa anterior se realizó la implementación para comprobar que lo que se esta proponiendo resuelve el problema de manera adecuada. Algunos elementos importantes de esta implementación son los siguientes:

  - Etiquetador de texto con topónimos
  - Generador automático de hechos
  - Interfaz de conexión CLIPS - Python
  - Consolidación de un gazetteer
  - Almacenamiento de la información
  - Visualización apropiada de los resultados

  Los elementos anteriores permitieron el desarrollo la herramienta de asignación de coordenadas geográficas como un producto final.

- **Experimentación**

  En esta etapa se llevaron a cabo una serie de experimentos para conocer si los resultados obtenidos garantizaron que el método propuesto fue el adecuado para el problema identificado. Para ello se tuvo que cumplir con las actividades siguientes:

- Pruebas: Se realizaron pruebas de la implementación y un análisis de los resultados.
- Ajuste de diseño: Permitió adaptar el diseño propuesto según los resultados obtenidos.

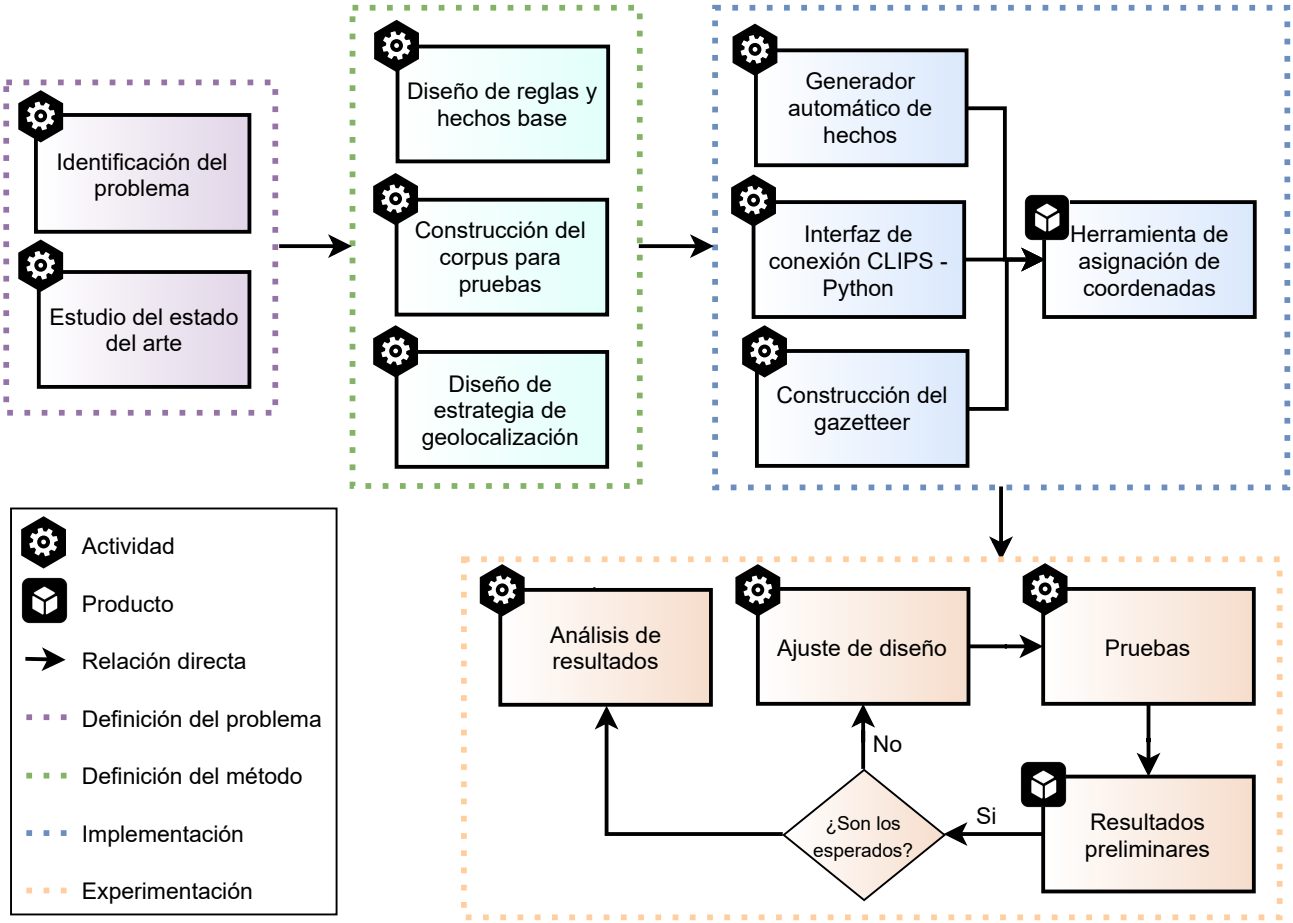


Figura 1.1: Metodología de desarrollo

## 1.5 Organización de la tesis

El documento está organizado de la siguiente manera. En el capítulo 2 se describen los conceptos básicos necesarios para el desarrollo de este trabajo de investigación, además de los trabajos del estado del arte con respecto a la desambiguación. En el capítulo 3 se presenta el método propuesto, describiendo los módulos que lo componen. En el capítulo 4 se presenta la evaluación experimental

y el análisis de los resultados obtenidos a partir de la implementación del método propuesto. En el capítulo 5 se presentan las conclusiones obtenidas, los inconvenientes que se presentaron durante la realización del trabajo y el trabajo futuro identificado.

# 2

# Marco teórico

En este capítulo se presentan los conceptos básicos del tema que constituyen el marco teórico. Asimismo se presentan algunos trabajos de diversos autores que constituyen el trabajo relacionado.

## 2.1 Introducción

Una de las tareas más importantes en Minería de Texto es la Recuperación de Información. Dentro de esta tarea se ubica el reconocimiento de objetos, personas, lugares y organizaciones, los cuales se denominan *entidades nombradas*. Es de especial interés para este trabajo el reconocimiento de entidades que representan topónimos.

Tal reconocimiento implica el uso de modelos computacionales que permiten la identificación automática de topónimos en un texto dado. Esta tarea de reconocimiento recibe el nombre de GNER (Geographical Named Entity Recognition).

Logrado lo anterior, es deseable conocer las propiedades geográficas (típicamente en términos de latitud y longitud) de los topónimos reconocidos, esta tarea es comúnmente denominada resolución

de topónimos (del Inglés toponym resolution)

Comúnmente, las tareas de GNER y resolución de topónimos se agrupan bajo el nombre de Geoparsing [2]. Uno de los desafíos más importantes en Geoparsing es resolver de manera automática posibles ambigüedades asociadas a homónimos, topónimos haciendo referencia a personas u otros tipos de entidades. La resolución de dichas ambigüedades es una de las motivaciones del trabajo presentado.

En las siguientes subsecciones se presentan los conceptos básicos para describir y presentar el método que aquí se propone.

## 2.2 Recuperación de información

La Recuperación de Información se refiere a encontrar documentos de naturaleza no estructurada, generalmente documentos de texto dentro de grandes colecciones [13]. La Recuperación de Información se encarga de recuperar los documentos relevantes para una consulta del usuario dejando fuera la mayor cantidad de documentos no relevantes. Es por esto que se considera que los documentos pueden ser inexactos y es probable que pequeños errores pasen desapercibidos, a diferencia de la recuperación de datos, en la que se espera que el resultado devuelto sea exacto [6].

Un sistema de Recuperación de Información trabaja con texto en lenguaje natural que no siempre se encuentra bien estructurado y podría ser semánticamente ambiguo. Entre las funciones de IR se encuentran el modelado, clasificación y categorización de documentos, arquitectura de sistemas, interfaces de usuario, visualización de datos, filtrado, idiomas, etc. La Recuperación de Información se enfoca en recuperar documentos basados en el contenido de sus componentes no estructurados. Una consulta puede especificar las características deseadas de los componentes estructurados y no estructurados de los documentos que se recuperarán.

En la Figura 2.1 se ilustra la arquitectura general de un sistema de Recuperación de Información. Esta arquitectura se basa en varios componentes que interactúan entre ellos: 1) un corpus, que es

un conjunto de documentos de texto debidamente recopilados [51], 2) un módulo de representación lógica que permite definir dichos documentos en términos sintácticos o semánticos, 3) un módulo de indexación que lleva a cabo la construcción de una estructura lógica, denominada índice, encargado de dar soporte a las búsquedas eficientes [42], 4) un módulo de búsqueda que acepta como entrada la consulta de un usuario, verifica el índice e identifica que documentos satisfacen dicha consulta [8], 5) un algoritmo de ranking encargado de determinar la importancia o relevancia de cada documento encontrado y retornar una lista [32] y 6) una interfaz de usuario que permite que se especifique la consulta mediante una expresión y se utiliza además para visualizar las respuestas del sistema.

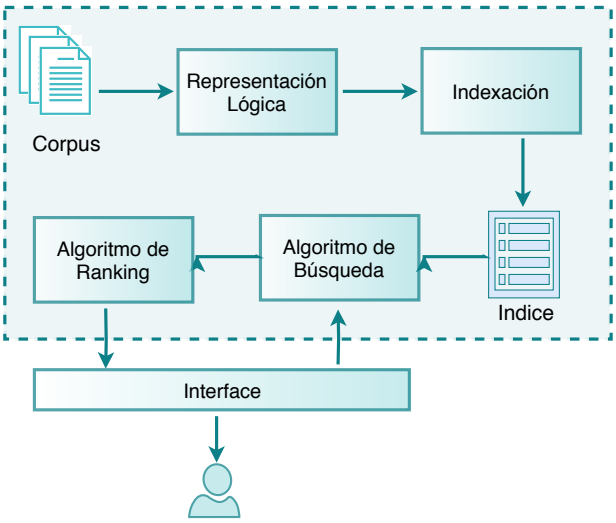


Figura 2.1: Arquitectura de un Sistema de Recuperación de Información [6]

Existen diversos modelos de recuperación populares. Estos modelos se pueden clasificar en cuatro categorías principales: modelos basados en conjuntos, modelos geométricos o algebraicos, modelos probabilísticos y modelos basados en aprendizaje automático [55].

### 2.2.1 Tipo de información

El diseño y desempeño de un sistema de Recuperación de Información está estrechamente relacionado con el tipo de información contenida en los documentos que almacena. Esta información puede ser de tres tipos diferentes:

- **Estructurada:** Un documento es estructurado si consiste en componentes y secciones organizados de acuerdo a una sintaxis bien definida. Comúnmente este tipo de documentos contienen información en estructuras tipo tabla, similares a una base de datos relacional; cuenta con múltiples tipos de registros, de manera que todos los registros de un tipo dado tengan una misma sintaxis [19].

- **Semi-estructurada:** Son documentos de texto que pueden compartir una estructura y una semántica común. De manera que una parte de los documentos se apegue a una estructura definida (por ejemplo, tablas) y otra parte es texto libre que cumple con una semántica específica.

- **No estructurada:** Se entiende como una colección de documentos en lenguaje natural sin una posición sintáctica bien definida en la que un motor de búsqueda pueda encontrar datos con una semántica dada [19]. No existe una sintaxis (externamente) bien definida para un documento determinado y mucho menos una sintaxis que compartan todos los documentos de una colección.

Los documentos para Recuperación de Información comúnmente están parcialmente estructurados, pueden tener un encabezado estructurado y un cuerpo no estructurado. Dentro del encabezado se encuentra generalmente una capa de meta-datos que consiste en información extra acerca del documento pero no contenido del mismo. En un documento bibliográfico estos meta-datos pueden estar conformados por un autor, título, editor, fecha de publicación, tema, resumen, números de catálogo, etc [19].

### 2.2.2 Recuperación de Información Geográfica

Comúnmente al escribir, las personas utilizan topónimos para describir ciertas situaciones; por ejemplo, dónde están, dar instrucciones de navegación, informar la ubicación de cierto evento, en general para transmitir información espacial que se basa en conocimiento compartido de estos

topónimos [53]. Los topónimos aparecen con frecuencia en documentos de texto con contexto geográfico. Gran parte de la información disponible en la Web está relacionada con topónimos y los usuarios incluyen el contexto geográfico en sus consultas a los motores de búsqueda web. La información geográfica se encuentra, por regla general, dentro del contenido de los documentos, se puede encontrar en forma de topónimos, direcciones, códigos postales, entre otras [52].

Procesar y recuperar documentos que contienen topónimos involucra tareas adicionales a un sistema de recuperación de información convencional. Estas tareas van desde identificar las palabras que denotan a los topónimos hasta la asignación de referencias geográficas. A continuación se describen cuatro técnicas para facilitar la recuperación de información geográfica. Estas tareas están asociadas a varias etapas del flujo de un motor de búsqueda:

- *Indexación geográfica:* Los topónimos son asociados a documentos mediante un índice espacial/geográfico para admitir operaciones de recuperación de información. La indexación geográfica puede ayudar a encontrar ubicaciones ideales para documentos en una infraestructura de almacenamiento distribuido, considerando que los usuarios de alguna región geográfica tienden a concentrar su interés en documentos de alcance global [29].

- *Expansión de la consulta:* Las búsquedas realizadas mediante topónimos pueden ser expandidas a través de los nombres relacionados, es decir, los topónimos que pertenecen a una misma jerarquía de subdivisión territorial. Por ejemplo, la búsqueda de documentos relacionados con un estado puede devolver documentos relacionados con cualquiera de sus municipios [5, 52].

- *Uso de topónimos en las consultas:* Cada vez más los topónimos son considerados como parte de las consultas debido a que se ha comprobado que conducen a mejores resultados [10, 16].

- *Clasificación geográfica:* Los motores de búsqueda web reconocen la posición geográfica del usuario y esto es considerado en la calificación de resultados, a través de esto se puede determinar qué tan cerca está un resultado de las intenciones expresadas por el usuario en

función de la naturaleza geográfica del término que se busca y del historial de búsqueda del usuario [2, 3].

En la Figura 2.2 se observa la arquitectura general de un sistema de recuperación de información geográfica, esta consta de 4 componentes descritos a continuación [24]: 1) una base de datos con información geográfica que puede provenir de fuentes de datos estructuradas y no estructuradas, 2) un módulo de indexación que permite construir una estructura de datos específica para facilitar la búsqueda, 3) un algoritmo de ranking que devuelve los resultados de la búsqueda de acuerdo con un puntaje asignado a cada documento con respecto a la búsqueda realizada, y 4) una interfaz permite al usuario ingresar consultas y recibir una respuesta del sistema.

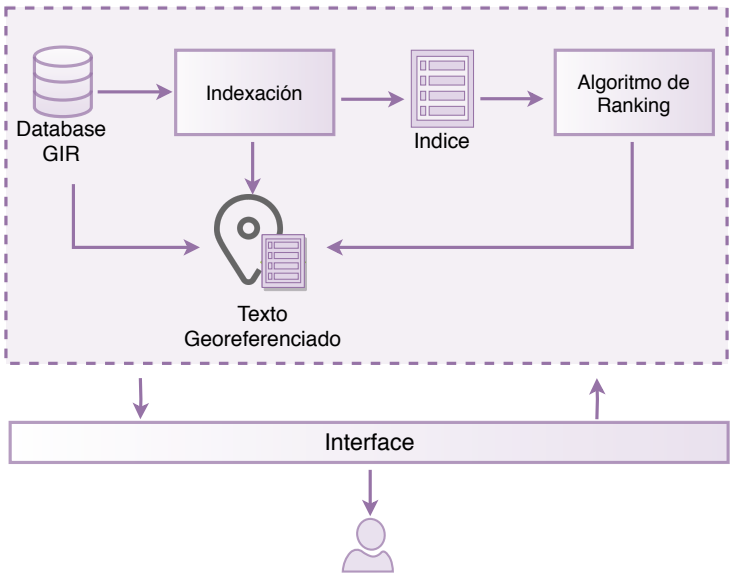


Figura 2.2: Arquitectura de un Sistema de Recuperación de Información Geográfica

## 2.3 Geoparsing

En geoparsing los topónimos contienen información geográfica. El proceso de geoparsing consta de dos partes: 1) Geotagging: consiste en identificar los topónimos presentes en el texto y sugerir

coordenadas posibles para el topónimo, y 2) Geocoding: que consiste que elegir la coordenada correcta de las posibles brindadas por el módulo anterior [20].

Geotagging es considerado un caso especial de reconocimiento de entidades nombradas (explicado en la siguiente sección) que es un problema común en procesamiento de lenguaje natural (NLP, por sus siglas del Inglés). La diferencia entre un reconocedor de entidades geográficas (GNER por sus siglas del Inglés) y un reconocedor de entidades nombradas es que el reconocedor de entidades geográficas solo recupera ubicaciones (no incluye personas, organizaciones, etc.), además de proporcionar coordenadas geográficas para cada una de ellas, tarea que un reconocedor de entidades nombradas no realiza.

Geocoding desde el punto de vista de NLP es el que considera la desambiguación de topónimos. Este trabajo se centra específicamente en esta tarea. Dada una lista de coordenadas candidatas para cada topónimo el objetivo es seleccionar la coordenada correcta, es decir, eliminar la ambigüedad. Finalmente, cada topónimo se vincula a un registro en una base de conocimiento geográfica como GeoNames.

## 2.4 Reconocimiento de Entidades Nombradas

El reconocimiento de entidades nombradas (NER, por sus siglas del Inglés) es una importante tarea de los sistemas de recuperación de información [47]. Se encarga de localizar y categorizar cadenas de símbolos con un significado especial dentro de un texto (nombres propios de personas, países, organizaciones, eventos, etc). Por ejemplo, de una noticia se pueden extraer los nombres de personas, organizaciones y ubicaciones [48]. Las entidades nombradas son palabras o frases que se nombran o clasifican en un tema determinado. Comúnmente contienen información clave en una oración que son considerados como objetivos para la mayoría de sistemas de procesamiento de lenguaje. El reconocimiento de entidades nombradas puede ser utilizado como una fuente de información para diferentes aplicaciones de Procesamiento de Lenguaje Natural [28].

El alcance de NER ha evolucionado en las últimas dos décadas. Originalmente se limitaba

a la extracción de nombres propios relacionados con las noticias, como nombres de personas, organizaciones y ubicaciones. Con la expansión del Procesamiento del Lenguaje Natural a otros dominios, esas pocas clases de entidades nombradas no fueron suficientes. NER se compone de dos tareas: 1) reconocimiento de límites en entidades nombradas y 2) reconocimiento de categorías en entidades nombradas. Para tener un sistema NER robusto para cualquier dominio dado es necesario de un corpus y léxicos etiquetados [27, 55]. Lo anterior significa que se debe contar de antemano con una lista de los nombres propios de un dominio de conocimiento que den indicios para identificar a otros del mismo tipo; y que se tenga del dominio de conocimiento un conjunto de documentos que contenga una parte de la lista de los nombres propios. El reconocimiento de topónimos se ubica dentro de la tarea general de reconocimiento de entidades nombradas, no añade más complejidad.

Los primeros enfoques NER estuvieron basados en reglas. Comúnmente se trata de tres reglas principales: 1) un conjunto de reglas de extracción de entidades nombradas, 2) un gazetteer para diferentes tipos de entidades nombradas y 3) un motor de extracción que aplica las reglas y los léxicos al texto [55]. Un nuevo enfoque son los métodos basados en datos y estadísticos [33]. El reconocimiento de entidades nombradas generalmente usa dos componentes principales:

1. *Datos de entrenamiento etiquetados*: consiste en un corpus donde se almacenan las entidades nombradas,

2. *Modelo estadístico*: consiste en una representación probabilística de los datos de entrenamiento.

Los sistemas NER se evalúan ejecutándolos sobre datos etiquetados por humanos y comparándolos con un corpus *gold-standard*. Un corpus *gold-standard* es un conjunto de documentos donde ya se tienen identificadas las entidades nombradas indicando el tipo de entidad. Esta comparación generalmente se encuentra a nivel de frase, dando crédito por coincidencias completas de categorías y sin crédito por coincidencias parciales [18]. Las métricas de evaluación comúnmente utilizadas son precisión y la cobertura. La precisión mide el porcentaje de entidades nombradas etiquetadas que coinciden con el *gold-standard* y que un sistema NER pudo reconocer. La cobertura mide el porcentaje

de entidades nombradas de un corpus *gold-standard* que un sistema NER puede reconocer. Existe una tercera medida, F-measure, esta es utilizada para combinar las dos métricas anteriores como se muestra a continuación [55]:

$$\text{Precision} = \frac{C}{L} \quad (1)$$

$$\text{Recall} = \frac{C}{G} \quad (2)$$

$$\text{F-measure} = \frac{2 \cdot Precision \dot{R}ecall}{Precision + Recall} \quad (3)$$

donde $L$ es el número de entidades nombradas etiquetadas, $G$ es el número de entidades nombradas en el corpus *gold-standard*, y $C$ es el número de entidades nombradas correctamente etiquetadas.

## 2.5 Ambigüedad

La ambigüedad en el sentido de las palabras se considera un problema debido a que una misma palabra puede reflejar distintas cosas. Sobre las ambigüedades del lenguaje, las computadoras necesitan procesar información textual no estructurada y transformarlas en estructuras de datos que deben analizarse para determinar el significado subyacente. La identificación computacional del significado de las palabras en un contexto determinado se denomina desambiguación del sentido de las palabras (WSD, por sus siglas del Inglés) [46]. Se considera un problema complejo de Inteligencia Artificial, es decir, una tarea cuya solución es al menos tan difícil como los problemas más difíciles en Inteligencia Artificial [38].

En el caso de topónimos también puede existir ambigüedad, esto debido a que los topónimos pueden compartir un mismo nombre con otro topónimo. La ambigüedad se origina debido a que los nombres de los topónimos se eligen usando palabras del lenguaje común o nombres propios. Algunos autores dividen esta ambigüedad en distintos tipos [4, 11, 56]:

- *Geo/Geo ambiguity*: Ocurre cuando un topónimo hace referencia a múltiples ubicaciones.

- *Geo/Non-Geo ambiguity*: Ocurre cuando un topónimo comparte el mismo nombre con algo que no es un topónimo, por ejemplo, con el nombre de una persona o una palabra en común.

- *Reference ambiguity:* Ocurre cuando un mismo topónimo está asociado a muchos nombres.

La ambigüedad provoca que la resolución de referencias a topónimos esté basada en el contexto de las palabras del texto. Un recurso importante para abordar la desambiguación es la determinación del alcance geográfico del documento, es decir, el conjunto de topónimos referenciados y relevantes para el contenido del mismo [35].

## 2.6 Georreferenciación

Esta tarea consiste en la asignación de coordenadas geográficas a puntos de interés en un mapa. Estos puntos de interés pueden ser cualquier elemento al que asignarle las coordenadas geográficas; por ejemplo ciudades, municipios, monumentos, ríos, parques, parajes, restaurantes y en general cualquier tipo de superficie o sus puntos. Es importante destacar que esta tarea no es exclusiva de recuperación de información. La georreferenciación resuelve dos grandes cuestiones simultáneamente [31]:

- Conocer la forma, dimensión y ubicación de cualquier parte de la superficie terrestre o cualquier objeto sobre ella.

- Vincular información espacial proveniente de distintas fuentes y épocas, necesarias para el desarrollo de los sistemas de información territoriales o geográficos.

Las referencias geográficas a topónimos pueden ser de distintos tipos: sencillas e inequívocas como coordenadas geográficas. Las fuentes de información de ubicaciones geográficas pueden ser estructuradas, como direcciones postales, o no estructuradas, como descripciones de topónimos. También pueden ser directas (como topónimos) o indirectas (como los que hacen referencia a

características culturales asociadas a topónimos), o explícitas o implícitas [35]. Es por esto que se necesitan herramientas o recursos que nos permitan obtener esos elementos para hacer la asociación entre el topónimo y las coordenadas que representa. A continuación se presentan algunos de estos recursos:

- **Gazetteer:** Consiste en información estructurada acerca de topónimos, vincula los topónimos a conceptos temáticos y a la representación de su extensión espacial, es decir, una huella geográfica (footprint) [17]. Un gazetteer es considerado como un diccionario geoespacial de nombres geográficos, es usado comúnmente para la recuperación de información indirectamente georreferenciada, es decir, basada en topónimos. Se encarga de vincular topónimos con coordenadas y datos temáticos, facilita la integración de información en línea de sistemas, y manejan eficientemente cantidades muy grandes de topónimos [37, 53].

  Debido a los desarrollos en los últimos diez años, existe un acuerdo sobre la estructura que debe tener un gazetteer. Los elementos mínimos requeridos se representan por una tupla $N,F,T$ donde $N$ es al menos un topónimo, $F$ es al menos una representación de la ubicación geográfica de acuerdo con un marco matemático (footprint) y $T$ es al menos un tipo (categoría, clase). Debido a que los topónimos presentan cambios de distintos tipos con el paso del tiempo, también es necesario incorporar fechas de modificación [17]. Cabe destacar que para esta propuesta es necesario de un elemento más, por lo que la representación de la tupla utilizada es $N,F,T,J$, donde $J$ es el nivel jerárquico al que pertenece el topónimo. Los gazetteers se generan de manera manual o semi-manual, pero su información siempre será corroborada y revisada por un especialista humano.

- **Ontología:** Es definida como una especificación formal explícita de un vocabulario de representación para un dominio compartido del discurso (definiciones de clases, relaciones, funciones y otros objetos) [21]. Una ontología comprende relaciones topológicas y de proximidad entre topónimos y datos de coordenadas dispersas que se encuentran en diversos gazetteers que

representan la huella geográfica de los topónimos. Conceptualmente, el nivel de abstracción de una ontología es lo suficientemente genérico como para ser utilizado y refinado por múltiples diccionarios geográficos y lo suficientemente detallado como para permitir la búsqueda directa de tipos [53]. Está compuesta por tres componentes principales: 1) una jerarquía de conceptos que captura todos los conceptos en un dominio dado, 2) los atributos de los conceptos y 3) una jerarquía de relaciones que presentan la relación no jerárquica entre los conceptos [54]. Existen dos enfoques distintos en la construcción de una ontología. Al igual que los gazetteers, las ontologías se generan de manera manual o semi-manual, necesitan la revisión de un especialista humano para que contenga datos e información correcta.

## 2.7 Trabajo relacionado

En esta sección se presenta una revisión de la literatura sobre el tema del trabajo de tesis. Además se incluye una comparación entre estos trabajos, tomando en cuenta algunos aspectos comunes. Finalmente se incluye una tabla que resume dichos trabajos.

Rupp *et al.* [45] presentaron una propuesta de desambiguación de topónimos donde su trabajo lo enfocaron en el contexto histórico, específicamente en un corpus de literatura de viajes del Distrito de los Lagos, Inglaterra. La idea principal fue la extracción de temas importantes de un texto y la ubicación en un mapa. Cabe destacar que el tipo de textos utilizados comprende desde el año 1622 hasta 1900, predominantemente desde los siglos XVIII y XIX. Realizaron la transcripción de 81 documentos, esto permitió la generación de un subcorpus de 28 textos, compuesto por topónimos y nombres de personas. Cada uno de los textos fue traducido del original de forma manual para evitar las inexactitudes que pudieran ocurrir al utilizar un proceso de OCR (Optical Character Recognition) y posteriormente se pasó por un traductor. Los autores utilizaron un sistema llamado VARD para corrección ortográfica. Este es un detector de variantes basado en un diccionario de formas estándar del inglés actual, lo que permite hacer una normalización en los topónimos mencionados. En cuanto

a las herramientas hicieron uso de un gazetteer histórico de Cumbria, además de una lista de nombres específicos de la región de Cumbria y una encuesta de ordenanzas. También hicieron uso de Edinburgh Unlock Geoparser [1], el cual ha sido aplicado anteriormente a textos históricos. El objetivo de este es identificar los topónimos mencionados en los textos. Una de las limitaciones del trabajo desarrollado es que al tratarse de textos históricos los topónimos actuales pueden no coincidir con los de los siglos pasados, por lo que la precisión varía, además de que al ser textos traducidos y transcritos la variación en la ortografía es otro punto a tener en cuenta.

A diferencia del trabajo anterior, en el enfoque propuesto por Tobin *et al.* [50] se contó con tres colecciones históricas digitalizadas de antemano, por lo que el problema de la variación en la ortografía se asume que ya está resuelto. Utilizaron técnicas de Extracción de Información para identificar los topónimos del corpus utilizado, hicieron uso de diferentes gazetteers, comparando los resultados obtenidos con anotaciones humanas de las tres colecciones. El proceso se aplicó en el corpus SpatialML [30], este es un corpus geoanotado de textos de periódicos. Se consideraron dos partes principales en este método, *geotagger* y *georesolver*. El primero procesa un texto de entrada e identifica las cadenas dentro de él que denotan topónimos. El segundo toma el conjunto de topónimos reconocidos como entrada, los busca en uno de los distintos tipos de diccionario geográfico y determina para cada topónimo cuál de los referentes posibles es el correcto. Para el caso de la ambigüedad clasifica las entradas candidatas en orden de probabilidad de acuerdo al contexto, si hay topónimos repetidos se basan en población o tipo de topónimo para tomar una decisión. Entre las desventajas de la propuesta los autores mencionan el hecho de que no encuentra topónimos si estos presentan algún problema en su escritura, es decir, si tiene alguna letra cambiada por otra no logra identificar de qué ciudad se trata.

Por otro lado, el trabajo desarrollado por Martins y Silva [32] consistió en una adaptación del algoritmo de ranking web de PageRank [39] a textos en Portugués para asignar documentos con alcance geográfico. Hay que destacar que PageRank es un método para calificar páginas web de forma

[1] http://www.ltg.ed.ac.uk/software/geoparser

objetiva y mecánica, midiendo efectivamente el interés humano y la atención que se les dedica [39]. Esta ponderación se considera de acuerdo con la frecuencia de ocurrencia de topónimos en el texto. Otra característica importante a tomar en cuenta es la utilización de referencias geográficas extraídas del texto y una técnica basada en ontologías. Dicha técnica consiste en dos ontologías geográficas que proporcionan tanto el vocabulario como las relaciones entre los conceptos geográficos; la primera de ella es multilingüe global y la segunda basada solo en la región de Portugal. Una de las limitaciones que tiene esta implementación es que el algoritmo de PageRank original otorga un mismo peso a todos los bordes (hipervínculos), lo que provoca que los nodos con más in-links (enlaces a otros sitios) tiendan a obtener rangos más altos, sean o no importantes para el problema, además de que no considera la etapa de Extracción de Información, por lo que se supone que los datos deben estar previamente preprocesados. Los autores reconocen que existen dos tipos de ambigüedad, la que ocurre cuando un mismo nombre se puede utilizar para varios topónimos y la que ocurre cuando un mismo topónimo puede tener múltiples nombres, a las que llaman *referent ambiguity* y *reference ambiguity*, respectivamente. Asimismo mencionan que la ontología y el software de extracción solucionan este problema pero no de qué manera.

Asimismo, Silva *et al*. [49] presentaron una propuesta en donde asignaron ámbitos geográficos a documentos en Portugués utilizando un algoritmo de clasificación de gráficos similar a PageRank. Además de que como un primer paso se enfocaron en la extracción de características, al reconocer y desambiguar las referencias geográficas. El método presentado hizo un uso extensivo de una ontología de conceptos geográficos e incluye una arquitectura de sistema para extraer información geográfica de grandes colecciones de documentos web. Cabe destacar que un ámbito geográfico se especifica como una relación entre una entidad en el dominio web (una página HTML o un sitio web) y una entidad en el dominio geográfico (como una ubicación o región administrativa). El alcance geográfico de una entidad web tiene la misma huella que la entidad geográfica asociada; este alcance asignado a un documento se otorga debido a la frecuencia de aparición de un término y además se toma en cuenta la similitud con otros documentos. Lo anterior se refiere a que si un topónimo se menciona

en el documento, el alcance del mismo está relacionado con la región mencionada. Por ejemplo, si el texto contiene "la Ciudad de Oporto", el alcance de esta página está relacionado con la Región de Oporto, además si un mismo topónimo se menciona en varias ocasiones es más probable que sea importante por lo que se puede concluir con mayor confianza que el alcance de ese documento corresponde a dicho topónimo. La asignación de alcances a documentos web se hace a través de un motor de búsqueda web (tumba.pt). El proceso está dividido en dos etapas: la primera identifica topónimos en los documentos web, otorgando un peso a cada uno de acuerdo con la frecuencia de aparición y a heurísticas HTML; la segunda etapa asigna un alcance final a cada documento web basado en las referencias geográficas encontradas y a los pesos asignados.

Por otro lado, Radke *et al.* [43] propusieron un algoritmo para el etiquetado geográfico de documentos web en Inglés considerando todos los topónimos juntos sin desambiguarlos individualmente. Para esta propuesta se requirió de un enfoque jerárquico y uno geométrico con la aplicación de una técnica de heurística, que aplica a grandes conjuntos de documentos. La heurística se basa en la observación que cuando se mencionan múltiples topónimos en un documento, la región más pequeña es la que abarca a todos desambiguando cada uno de ellos. El método requiere de un gazetteer y de una herramienta NER para su procesamiento; este enfoque está pensado para texto plano, específicamente documentos web. Consiste en tres pasos: el primero es jerárquico, en el cual se le da un enfoque a nivel país al documento, el segundo es a nivel geométrico, en el cual se realiza un tipo de desambiguación y define el alcance del documento, el tercer paso es el que reporta el alcance final del documento. Como resultado final los autores reportaron latitud y longitud de los topónimos mencionados en los textos.

A su vez, Nesi *et al.* [40] presentaron un sistema llamado GeLo, el cual extrae direcciones y coordenadas geográficas de empresas comerciales, instituciones y otras organizaciones de dominios web. Utilizaron dos conjuntos de datos, el primero compuesto por 6 millones de URLs y el segundo compuesto de 100,000 URLs. Este proceso de extracción se basó en técnicas de NLP, específicamente Part-Of-Speech-Tagging, reconocimiento de patrones y anotaciones. Las pruebas

realizadas se centraron en dominios web de organizaciones ubicadas en la región de Toscana, en Italia. La plataforma está desarrollada para el idioma Italiano e Inglés, además de una modalidad independiente del idioma. Basó su arquitectura en dos módulos, el primero es una herramienta de rastreo para la indexación de documentos, el segundo es un analizador lingüístico que toma como entrada los documentos y las páginas recuperadas en las URLs de la web obtenidas por el módulo anterior. La arquitectura de su sistema consiste tres módulos: el primero es un crawler, que permite buscar grandes cantidades tanto de textos como de documentos específicos, el segundo está dedicado a la extracción de direcciones compuesto por un analizador lingüístico basado en NLP que cuenta con dos modalidades, dependiente e independiente del idioma, y finalmente un módulo de geocodificación encargado de recuperar las coordenadas de las direcciones extraídas.

Por otro lado Inkpen *et al.* [26] desarrollaron un algoritmo que extrae expresiones compuestas de una o más palabras para cada topónimo, hicieron uso de un clasificador Conditional Random Fields (CRF), el cual se basa en un modelo gráfico no dirigido que se utiliza para predicciones no estructuradas [23]. Se enfocaron en topónimos presentes en tweets, para ello definieron reglas de desambiguación basadas en heurísticas. El corpus utilizado contiene tweets solo en el idioma Inglés, en mayor cantidad enfocado en los estados y provincias de los Estados Unidos y Canadá. El corpus fue recolectado por los mismos autores, haciendo distinción entre seis marcas de teléfonos celulares debido a que la API de Twitter para esas marcas permite filtrar tweets basados en el lenguaje, origen geográfico, entre otras características. La recolección se realizó de junio de 2013 a noviembre del mismo año, recopilando alrededor de 100 tweets por día, por lo que al finalizar la fecha se consiguió un total de 20 millones de tweets. Debido a que la cantidad de datos fue demasiada se utilizó una muestra al azar de 1000 tweets para cada subconjunto correspondiente a cada marca de teléfono, por lo que el corpus final consistió en 6000 tweets. Hicieron uso del gazetteer de GeoNames, el cual incluye información adicional a los nombres y coordenadas de los topónimos, por ejemplo, población, nivel administrativo, entre otros. El proceso de desambiguación se divide en cinco etapas: 1) recuperación de candidatos, 2) tipo de filtrado, 3) revisión de topónimos adyacentes, 4) verificación del contexto

global y 5) significado predeterminado.

Aunque algunos trabajos descritos anteriormente han sido desarrollados para múltiples idiomas (Martins y Silva [32] y Silva *et al.* [49]), con base en la revisión realizada, sólo un trabajo se centra en el Español mexicano [34]. Molina-Villegas *et al.* [34] presentaron un sistema que permite detectar entidades georreferenciadas de México en discurso libre, tanto oral como escrito. La desambiguación de topónimos en idiomas diferentes a Inglés, Portugués e Italiano es una importante área de oportunidad, ya que cada idioma tiene sus propios patrones de lenguaje, los cuales permiten obtener información adicional para identificar el nombre de las entidades a georreferenciar. El trabajo de Silva *et al.* [49] fue probado para Español, específicamente de topónimos de España. El trabajo de Martins y Silva [32] no especifica qué idiomas procesa, solo menciona que es posible usarlo en varios idiomas. Esto se debe a que no considera el contexto del texto y solo extrae cada topónimo por separado para georreferenciar. A pesar de esto las pruebas fueron realizadas con corpus en Portugués. En la edición 2002 de CoNLL [48] [2] se consideró por primera vez al Español como idioma de interés [47]. No obstante, los datos empleados en el CoNLL fueron recopilados por la Universidad Politécnica de Cataluña y la Universidad Autónoma de Barcelona, por lo que las anotaciones están enfocadas en documentos de España, dejando de lado cualquier otra variante, incluida la mexicana.

Los trabajos de Silva *et al.* [49], Inkpen *et al.* [26] y Tobin *et al.* [50] destacaron la necesidad de recursos similares pero para topónimos de otros países. En este sentido, el desafío principal fue desarrollar y obtener un corpus de calidad etiquetado para tales topónimos. Silva *et al.* [49] identificaron los posibles problemas de desambiguación y los resolvieron con base en la frecuencia de ocurrencia asociada al texto, es decir, analiza el discurso e identifica el tipo de topónimo posible (ciudad, pueblo, río, etc) con respecto a las palabras mencionadas antes o después del topónimo. Inkpen *et al.* [26] especificaron una serie de reglas que permiten desambiguar pero está enfocado en estados y provincias de Estados Unidos y Canadá, dejando de lado entidades más pequeñas como

[2] Grupo de Interés en Aprendizaje de Languaje Natural de la ACL (Asociación Americana de Linguistica Computacional)

barrios o ciudades, o incluso variaciones en los topónimos. Tobin *et al.* [50] clasificaron las entradas candidatas en orden de probabilidad según el contexto.

En la Tabla 2.1 se muestra un resumen sobre los trabajos mencionados anteriormente, resaltando los conceptos de mayor importancia. Se considera el tipo de entrada, el idioma sobre el que trabaja, la herramienta (que permite saber qué toman en cuenta para hacer la georreferenciación) y si consideran la desambiguación de los topónimos o no.

| **Autor** | **Año** | **Idioma** | **Herramientas** | **Desambiguación** | Grano Fino | Contexto |
|---|---|---|---|---|---|---|
| Molina-Villegas *et al.* [34] | 2019 | Español | Gazetteer | No | Si | No |
| Radke *et al.* [43] | 2018 | Inglés | Gazetteer | No | No | Si |
| Inkpen *et al.* [26] | 2017 | Inglés | Gazetteer | Si | No | Si |
| Nesi *et al.* [40] | 2014 | Italiano e Inglés | Ontología | No | No | No |
| Rupp *et al.* [45] | 2013 | Inglés | Gazetteer | No | Si | No |
| Tobin *et al.* [50] | 2010 | Inglés | Gazetteer | Si | No | Si |
| Silva *et al.* [49] | 2006 | Inglés, Portugués y Español | Ontología | Si | No | No |
| Martins y Silva [32] | 2005 | Inglés y Portugués | Ontología | No | No | No |

Tabla 2.1: Categorización de métodos NER Geográficos según el tipo de entrada, idioma, herramientas externas y método de desambiguación.

# 3

# Solución propuesta

En este capítulo se describe el método propuesto para resolver la ambigüedad de topónimos presentes en un texto en Español. Primero se da una descripción general del método, posteriormente se presenta una descripción detallada de cada una de las etapas que componen dicho método.

## 3.1 Descripción general

El método propuesto forma parte de un método mayor de geoparsing. El método de geoparsing consiste de dos etapas principales: a) un reconocedor de entidades geográficas (GNER) y b) un desambiguador de topónimos.

Es importar destacar que el método GNER no se desarrolló como parte de esta tesis. Sin embargo, se hace referencia a él debido a que es parte del problema de geoparsing descrito anteriormente en el Capítulo 1. El objetivo de este método es reconocer cada topónimo mencionado en un texto en Español e identificarlos dentro del texto mediante etiquetas. Reconocer los topónimos en un texto permitirá inferir información geográfica contenida en el mismo.

La segunda parte (el método propuesto) permite desambiguar topónimos identificados en un texto, a partir de un conjunto de reglas y hechos. Esta parte el método de geoparsing se basa en la forma en que los humanos resolvemos la ambigüedad, mediante asociaciones con otros elementos del un mismo texto, identificando elementos dentro de contextos anidados. La propuesta está compuesta de dos módulos principales, además de otros elementos necesarios, éstos son ilustrados en la Figura 3.1 y descritos de manera general a continuación:

- **Reglas:** Se definió un conjunto de ocho reglas (definidas en la Tabla 3.3), que permiten realizar el proceso de desambiguación. Estas reglas están diseñadas tomando en cuenta el proceso en el que los humanos resuelven este tipo de ambigüedades.

- **Generador de hechos:** Es un módulo de software que permite generar, a partir del texto de entrada, una serie de hechos derivados del texto que activan alguna de las reglas previamente definidas. Cada uno de estos hechos se activa con base en el orden de los topónimos del texto, la información contenida de un gazetteer y al estado de una *pila de desambiguación*. En total, y hasta el momento, se cuentan con 6 hechos, descritos en la Tabla 3.1.

- **Pila de desambiguación:** Es una estructura de datos con la funcionalidad de una pila, en la que se va colocando cada topónimo analizado con sus respectivas propiedades geográficas (latitud, longitud, código de estado, entre otras).

- **Motor de inferencia:** Es un módulo encargado de identificar la regla o reglas a activar de acuerdo a los hechos generados a partir del contenido del texto.

- **Ejecutor de acciones:** Es un módulo que, como su nombre lo indica, ejecuta las acciones definidas para cada regla activada a partir de algún hecho.

- **Visualización:** Es un módulo que permite desplegar visualmente en un mapa el estado final de la pila, es decir, ilustra en un mapa cada topónimo presente en la *pila de desambiguación*.

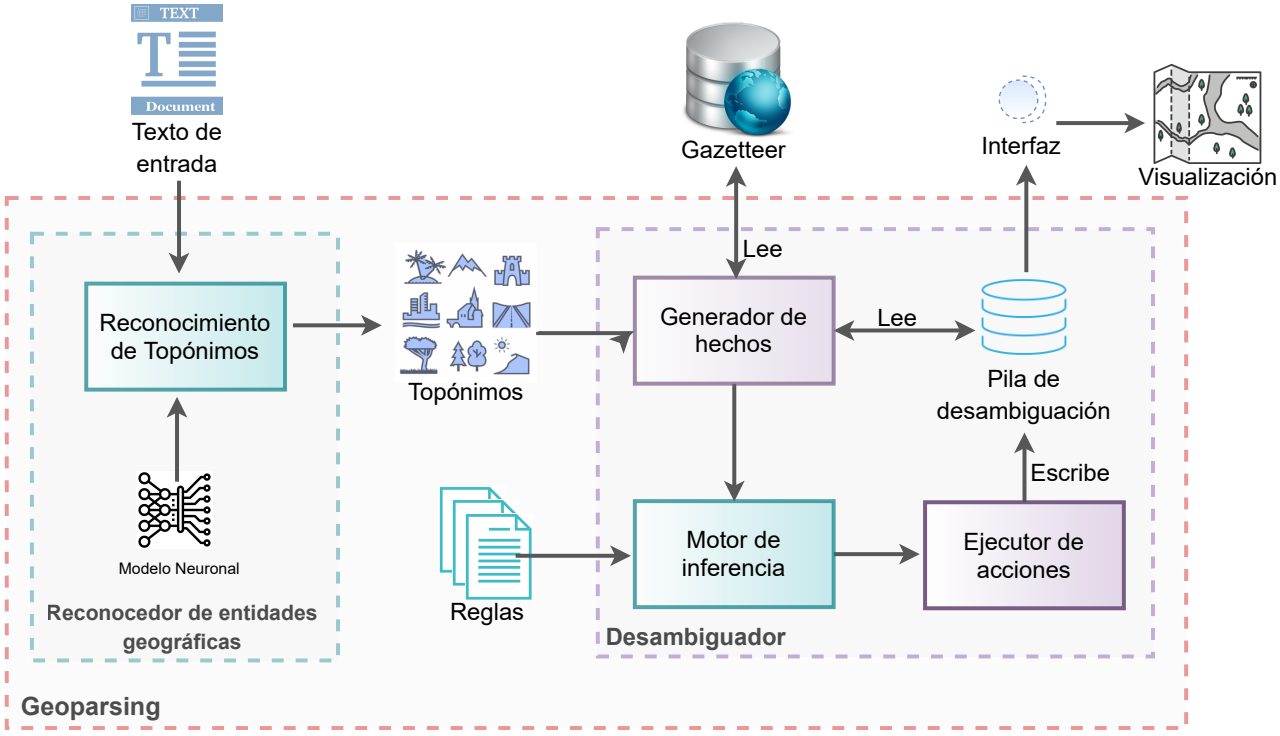


Figura 3.1: Propuesta de solución

## 3.2 Módulos del método

En esta sección se describen de manera detallada cada uno de los métodos que forman parte del método propuesto. En la Figura 3.1 puede verse la arquitectura del método propuesto.

- **Reconocedor de entidades geográficas**

  Este módulo es el encargado de reconocer todos los topónimos presentes en un texto de entrada. Para ello busca e identifica los topónimos en el texto. Los identifica mediante etiquetas, este etiquetado consiste en delimitar dentro del texto la(s) palabra(s) que compone un topónimo entre las etiquetas *<location>* y *</location>*

  Un ejemplo de este proceso se muestra a continuación, donde a partir de un texto original (A), el módulo identifica las menciones de topónimos, las cuales coloca entre etiquetas definidas (B).

A) La aparición de fábricas atrae a familias trabajadoras, primero del campo de **Nuevo León** y después, de estados vecinos, como **San Luis Potosí**, de ahí el nombre del famoso puente y barrio de **San Luisito**.

B) La aparición de fábricas atrae a familias trabajadoras, primero del campo de <*location*> ***Nuevo León*** </*location*> y después, de estados vecinos, como <*location*> ***San Luis Potosí*** </*location*>, de ahí el nombre del famoso puente y barrio de <*location*> ***San Luisito*** </*location*>.

La herramienta GNER que se utiliza para esta tarea se llama GeoparseMX [1]; este es un reconocedor de entidades geográficas basado en un modelo de red neuronal pre-entrenado con un corpus que incluye ubicaciones mexicanas llamado Corpus de Entidades Georreferenciadas de México (CEGEOMEX[2]). Este corpus tiene un total de 61,946 palabras, distribuidas en 1,233 documentos de diversos medios digitales en México, también contiene un conjunto de 5,870 entidades geográficas nombradas que han sido etiquetadas manualmente. Se hizo uso de este sistema debido a que permite reconocer topónimos en el contexto mexicano, esto es algo que se desea realizar con el trabajo desarrollado en esta tesis.

- **Desambiguador**

  Una vez reconocidos y etiquetados los topónimos en el texto, el paso siguiente es desambiguarlos y asignarles las propiedades geográficas correspondientes a cada uno. La forma común de realizar esta asignación es consultar en un gazetteer el topónimo de interés y asignar el nombre del topónimo con la primera opción devuelta. El problema de realizar una asignación de esta manera es que se asigna, sin más, la primera opción asumiendo que es el topónimo deseado, cuando realmente puede ser otra. Esto hará que se asigne información de un topónimo real a una etiqueta de otro topónimo diferente, lo cual generará ambigüedad cuando un usuario

[1] Todas las url's mencionadas se visitaron en 07/2020
http://geoparsing.geoint.mx/mx
[2] http://geoparsing.geoint.mx/mx/info

haga uso de esa información. La forma en la que se aborda el problema en el método propuesto consiste en tomar en cuenta todos los resultados devueltos por el gazetteer para cada topónimo y determinar cuál de las opciones es la más adecuada de acuerdo al contexto de las palabras que rodean la etiqueta del topónimo en cuestión. Es decir, se toman en cuenta las palabras y topónimos vecinos del topónimo en cuestión.

Para comprender mejor el problema de ambigüedad en la Figura 3.2 se muestra un ejemplo. En el texto presentado el topónimo ambiguo es *Tonalá*, ya que según lo descrito en el texto, representa un municipio del estado de Chiapas y también una calle en el estado de Jalisco.

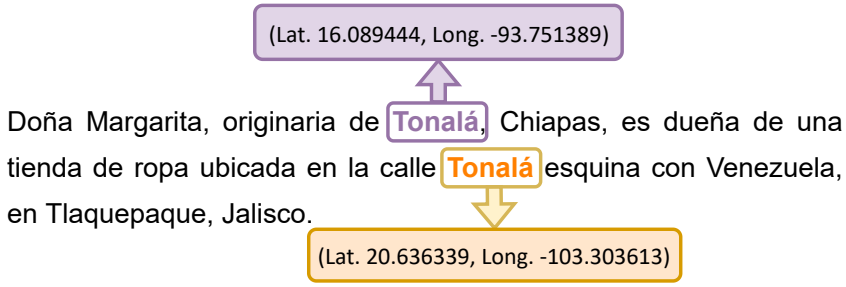


Figura 3.2: Ejemplo de ambigüedad en topónimos en un texto

El método de desambiguación está compuesto por tres módulos (como se ilustra en la Figura 3.1): un generador de hechos, un motor de inferencia y el módulo encargado de ejecutar las acciones indicadas por el motor de inferencia. Cada módulo se describe a continuación:

El **Generador de hechos** toma como entrada el texto etiquetado producido por el módulo Reconocedor de Entidades Geográficas y analiza cada topónimo haciendo consultas tanto al gazetteer como a la *pila de desambiguación*; esto con el objetivo de ir relacionando la información contenida en el gazetteer sobre el topónimo y su contexto en el texto (almacenado en la pila de desambiguación). El *gazetteer* contiene información sobre el topónimo, latitud y longitud, población, código de estado o de municipio, característica y código del área, elevación. La *pila de desambiguación* almacena cada topónimo encontrado, además de sus propiedades geográficas (latitud, longitud, código de estado y de municipio) del mismo, lo que permite

conocer las relaciones que existen entre cada uno de los topónimos presentes en texto. Se utiliza una pila adicional denominada *pila de conflictos*, ésta es encargada de almacenar cada topónimo que no sea posible georreferenciar al inicio del proceso. Es decir, en esta pila se colocan los topónimos que no se encuentran en el gazetteer, por lo que no es posible asignarles una coordenada. Cabe aclarar que esta pila es revisada al finalizar el proceso de desambiguación y cada topónimo que en ella permanezca es pasado a la *pila de desambiguación* asociándolo a otro previamente procesado. El topónimo al que se asignará depende de la acción indicada en la regla activada.

El **Motor de inferencia** es el encargado de activar o desactivar reglas para ir obteniendo conclusiones parciales sobre los topónimos a partir de los hechos que recibe. Si se cumplen con los hechos definidos para cierta regla ésta es activada. El conjunto de *reglas* permite realizar ciertas acciones sobre un topónimo para descartar los topónimos que no se mencionan en el texto pero que comparten un mismo nombre con el topónimo a procesar.

Finalmente el módulo **Ejecutor de acciones** permite aplicar la regla o reglas activadas por el motor de inferencia, ejecuta las acciones indicadas por la regla o reglas que se hayan activado y realiza una acción sobre la *pila de desambiguación*. Esta acción se encuentra definida en la regla activada, las posibles acciones a realizar se describen en la Tabla 3.1.

- **Hechos**

  Los hechos son una serie de sucesos o eventos que se derivan a partir de lo que se describe en el texto sobre los topónimos y la información de estos en el gazetteer. Permiten activar una o varias reglas, permiten realizar una acción sobre el topónimo en cuestión, y en algunos casos sobre el topónimo anterior. En la Tabla 3.1 se describen los seis hechos que se emplean en el método. Donde $\mathbb{G}$ representa el gazetteer, $\mathbb{S}$ representa la pila de desambiguación, $\mathbb{C}$ representa la pila de conflictos.

- **Reglas**

| Hecho | Descripción |
|---|---|
| $P_1$ | El topónimo existe en $\mathbb{G}$. |
| $P_2$ | La pila de desambiguación ($\mathbb{S}$) está vacía. |
| $P_3$ | El topónimo a procesar es predecesor del topónimo ya procesado. |
| $P_4$ | Existe una relación bidireccional entre el topónimo procesado y el topónimo a procesar. |
| $P_5$ | No existen más topónimos a procesar. |
| $P_6$ | La pila de conflictos $\mathbb{C}$ no está vacía. |

Tabla 3.1: Conjunto de hechos empleados en el método propuesto

Una regla está formada por dos partes, antecedente y consecuente. Antecedentes son los hechos identificados a partir del texto que permiten activar o desactivar una regla. Consecuentes son las acciones que se realizarán como resultado de que se activen o desactiven las reglas según el antecedente; estas acciones implican cambios en la pila de desambiguación y en la pila de conflictos. En la Tabla 3.3 se describen las reglas con sus respectivos antecedentes y consecuentes.

- **Categorías geográficas**

  Los topónimos presentes en el gazetteer cuentan con un campo que nos permite asignar un valor jerárquico a cada topónimo. En la Tabla 3.2 se listan etiquetas que identifican las categorías geográficas a las que puede asociarse un topónimo. A cada categoría geográfica se le asigna un valor que representa el nivel jerárquico. Más adelante esta categoría es la que se toma como clase para determinar si el método es capaz de reconocerla o no. La columna descripción contiene el orden administrativo de la región a la que se refiere el topónimo.

- **Visualización**

  Al terminar el proceso de desambiguación la pila de desambiguación mantiene todos los topónimos mencionados en el texto con sus respectivas propiedades geográficas. Esto permite que se puedan usar esos datos de diferentes maneras, por ejemplo, almacenarlos en alguna base de datos o visualizarlo, como lo hace el método propuesto. Para ello se desarrolló una

| Categoría geográfica | Valor | Descripción |
|---|---|---|
| ADM1 | 1 | Estado |
| PPLA | 2 | Capital |
| ADM2 | 3 | Municipio de nivel 1 |
| PPLA2 | 4 | Municipio de nivel 2 |
| PPL | 5 | Lugar poblado |
| LCTY | 6 | Localidad |
| OTROS | 7 | Lugar específico |
| NULL | 9 | No se encuentra en el gazetteer |

Tabla 3.2: Categorías geográficas, su respectivo valor y descripción.

interfaz que toma como entrada el contenido de la pila y lo representa en un mapa para mejor comprensión por parte del usuario final.

## 3.3 Algoritmo

El Algoritmo 1 muestra, en lo general, la secuencia de invocación de los procesos fundamentales del método de geoparsing en el que se ubica el método propuesto de desambiguación. La función `entity_recognition()` representa el módulo de reconocimiento de entidades geográficas, la función `rules_inference()` representa el módulo de desambiguación, ambos módulos fueron explicados anteriormente. La función `solve_conflicts()` permite realizar una revisión final del estado de las pilas de desambiguación y de conflictos para asegurarse que no quedan topónimos sin georreferenciar.

Para efectos ilustrativos, en el Apéndice B se puede encontrar un ejemplo detallado de la ejecución del Algoritmo 1, utilizando como entrada un documento del corpus que se utilizó durante la etapa de experimentación.

```
   Data:
   𝔻: Documento,
   𝔾: Gazetteer,
   Result: Toponym Resolution over 𝔻
 2 /* Inicialización de pila de desambiguación y pila de
      conflictos                                          */
 4 𝕊 ← ∅;
 6 ℂ ← ∅;
 8 /* Reconocedor de Entidades Geográficas               */
 9 𝕃=entity_recognition(𝔻);
11 /* Desambiguador de entidades geográficas             */
12 foreach e ∈ 𝕃 do
13 |   rules_inference(e,𝕊,𝔾);
14 |   if there is a conflict then
15 |   |   ℂ.push(e)
16 |   end
17 end
18 solve_conflicts(ℂ,𝕊);
19 return 𝕊
```

**Algoritmo 1:** Algoritmo de Geoparsing

## 3.4 Implementación

A partir de la definición del método de desambiguación de topónimos descrito en la sección anterior, se llevó a cabo su implementación para analizar resultados parciales y realizar las modificaciones necesarias para corroborar su desempeño. La implementación fue dirigida por los módulos descritos anteriormente. El diseño de los módulos de software se realizó aplicando el paradigma de programación orientado a objetos, dando origen al diagrama de clases descrito en el Apéndice 3.4.1. Las características particulares de la implementación se describen a continuación:

- Se usó Python 3.7.3[3] como lenguaje de programación. La generación de hechos y las pilas de desambiguación y de conflictos que se utilizaron están desarrolladas en este lenguaje.
- Se utilizó CLIPS 6.30[4] como motor de inferencia. CLIPS permite determinar qué reglas serán

[3] https://www.python.org
[4] http://www.clipsrules.net

las que se activarán dependiendo de los hechos que recibe a partir del texto (ver Apéndice A).

- Se utilizó Leaflet 1.6.0[5] para la visualización, ésta es una librería de JavaScript que permite hacer uso de mapas.

- Se usó Geonames[6] como gazetter. Para poder utilizarlo fue necesario pasarlo a una base de datos en MySQL 8.0.18[7].

- El entorno de programación utilizado para realizar el desarrollo fue Sublime Text 3.2.2[8]

- Para la visualización del mapa se usó Google Chrome 83.0.4103.116

En la Figura 3.3 se ilustra la interfaz de usuario. Del lado derecho se encuentra un cuadro de texto en el que se ingresa el texto que contiene los topónimos a desambiguar. Hay que recordar que el texto debe tener los topónimos identificados con las etiquetas <location>y </location>. Debajo de este se sitúa el botón que permite conocer el estado final de los topónimos (visualizándolo en el mapa). Del lado izquierdo se encuentra el mapa, en el que se ubica cada topónimo identificado en el texto.

[5]https://leafletjs.com
[6]https://www.geonames.org
[7]https://dev.mysql.com/downloads/workbench
[8]https://www.sublimetext.com/3

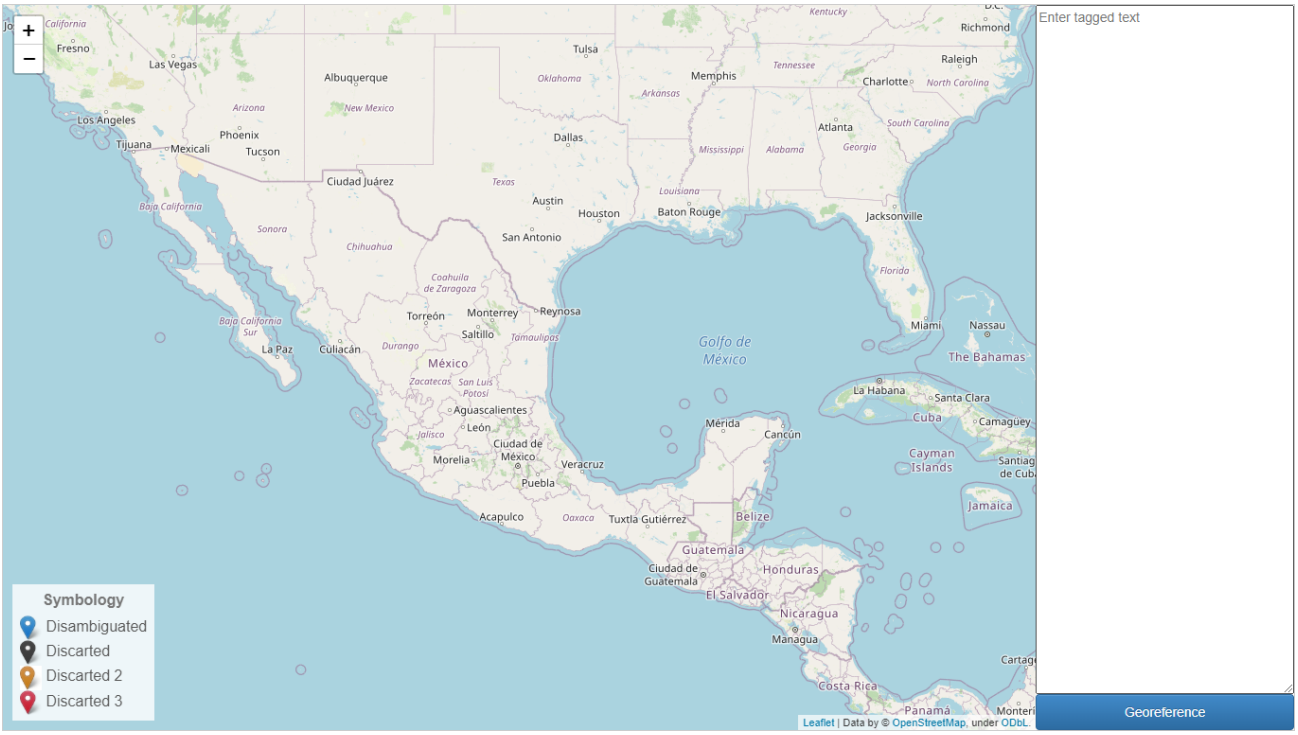


Figura 3.3: Interfaz de usuario

### 3.4.1 Diagrama UML de la implementación

El método propuesto se implementó siguiendo el paradigma orientado a objetos, esto da origen al diagrama UML presentado en la Figura 3.4. Las clases y métodos que componen este diagrama son descritos a continuación:

- *georeference* es la clase principal, dentro de esta se encuentran los métodos que permiten realizar las acciones indicadas por el motor de inferencia.

- *stack* es la pila de desambiguación.

- *factsGenerator* es la clase que genera los hechos a partir de una consulta al gazetteer y a la pila de desambiguación. A través de estas consultas activa o no ciertos hechos.

- *entity* es la clase que controla a la entidad, dentro de esta se encuentran cada una de las características necesarias para identificarla.

- *experimentation* es la clase creada para la etapa de experimentación, esta nos permite vaciar los resultados en una base de datos para realizar el análisis de los resultados al finalizar la desambiguación de todas las entidades encontradas en el corpus.

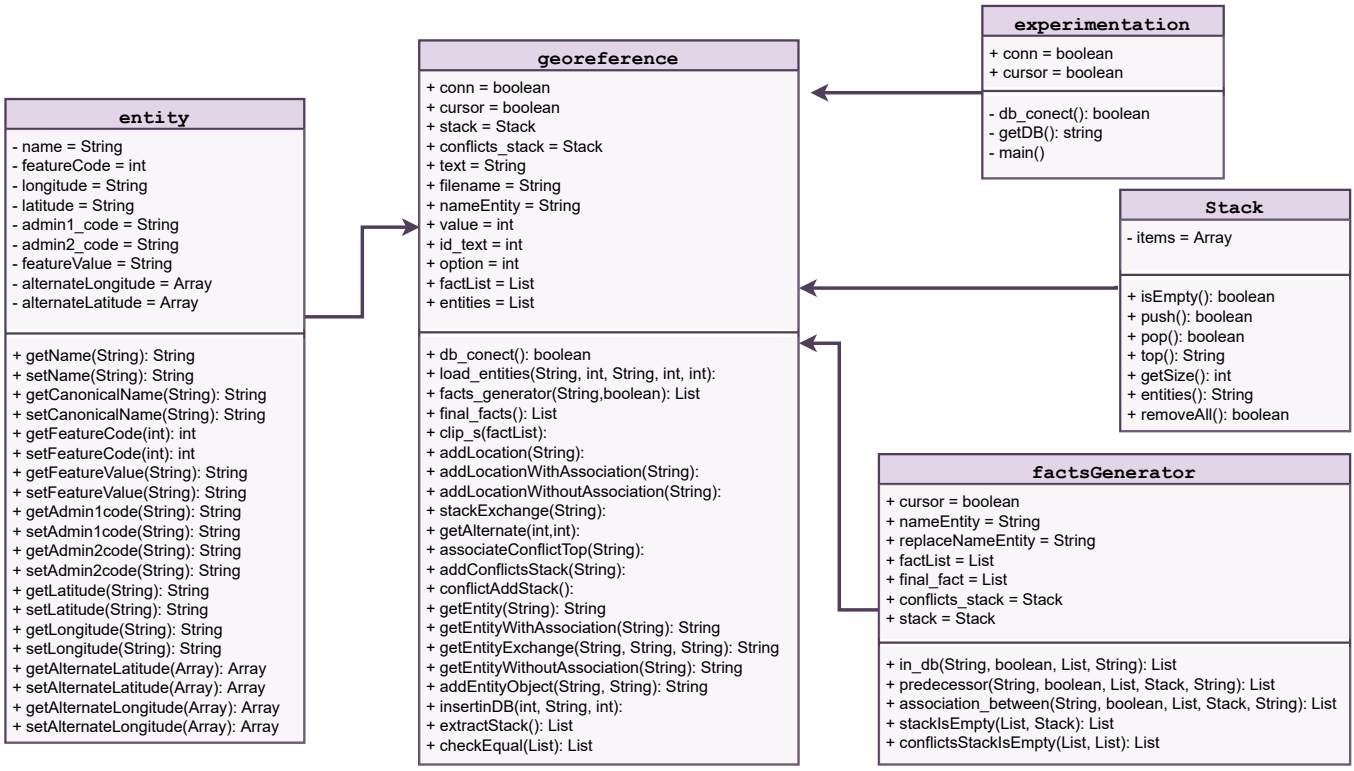


Figura 3.4: Diagrama UML

### 3.4.2 Ejemplo de aplicación

Para ilustrar mejor el funcionamiento del método propuesto, se presenta la Figura B.1, en la cual se muestra el texto inicial, las reglas activadas con cada topónimo ingresado y la pila de desambiguación con el estado después de procesar todos los topónimos.

En la parte superior de la imagen se encuentra el texto de entrada. Como se puede observar, cada topónimo se resalta en negritas y con ciertos colores. Estos colores se explican en el diagrama del lado izquierdo, en este se ejemplifica la jerarquía entre topónimos. Como se puede observar, de color morado se encuentra la entidad de más alto nivel jerárquico, en este caso, el país México, este se incluye en el diagrama a pesar de no expresarse directamente en el texto. Los demás grupos de colores representan un código distinto explicado en la Tabla 3.2.

Debajo del texto, del lado derecho se encuentra la lista de reglas activadas para cada topónimo presente en el texto. Las reglas definidas se describen en la Tabla 3.3. La activación de cualquier regla implica cambios sobre la pila de desambiguación. Esta pila se encuentra debajo de la lista de reglas, en ésta se muestra el estado final de la pila.

Las acciones siguientes representan un seguimiento paso a paso del comportamiento del método propuesto con respecto al ejemplo de la Figura B.1.

1. El topónimo *Mueblería Tu Hogar* no se encuentra en el gazetteer ($\mathbb{G}$), además debido a que la pila de desambiguación ($\mathbb{S}$) se encuentra vacía, se activa la regla $R_5$. La cual indica que el topónimo será colocado en la pila de conflictos ($\mathbb{C}$). La pila de conflictos es la encargada de almacenar los topónimos con coordenadas desconocidas hasta que es posible asociarlos con otros.

2. El topónimo *Ciudad de México* se encuentra en $\mathbb{G}$, además la pila $\mathbb{S}$ todavía está vacía. Esto provoca la activación de la regla $R_0$, lo cual lleva a colocar este topónimo en $\mathbb{S}$.

3. El topónimo *Azcapotzalco* se encuentra en $\mathbb{G}$, además $\mathbb{S}$ no se encuentra vacía, por lo que la regla $R_1$ es activada, lo cual indica que Azcapotzalco será colocado en el top de $\mathbb{S}$ asociada a la entidad anterior (*Ciudad de México*).

4. El topónimo *Ixcatán* se encuentra en $\mathbb{G}$, $\mathbb{S}$ no está vacía y este topónimo no está asociado al anterior. Estos hechos activan la regla $R_2$, la cual nos indica que *Ixcatán* será colocada sin asociación con las anteriores.

5. Los hechos activados por el siguiente topónimo, *Zapopan*, provocan la activación de la regla $R_3$. Esta regla realiza un cambio en la pila y permite colocar este nuevo topónimo asociado al antepenúltimo topónimo en $\mathbb{S}$ (*Azcapotzalco*) sin asociarla a *Ixcatán*. Hasta el momento en el top de la pila se encuentra *Ixcatán*, debajo de ésta *Azcapotzalco* y en la base la pila *Ciudad de México*.

6. Al igual que en el paso anterior, *Jalisco* activa la regla $R_3$, colocando nuevamente *Ixcatan* en el top y *Jalisco* debajo de ésta, pero en esta ocasión *Ixcatan* y *Jalisco* si están asociadas.

7. El topónimo *Pedregal de Santo Domingo* activa la regla $R_6$, lo cual indica que este topónimo será colocado en el top de $\mathbb{S}$ sin relación con el topónimo anterior.

8. El topónimo *Coyoacán* activa la regla $R_3$, lo que realiza un cambio en la pila, colocando este topónimo debajo del top actual.

9. El topónimo *Zapatería Juárez* no se encuentra en $\mathbb{G}$, lo que activa la regla $R_4$.

10. Como paso final se hace una revisión de $\mathbb{C}$ para conocer su estado, como en este caso *Mueblería tu hogar* se encuentra ahí, se activa la regla $R_7$.

En la Figura B.2 se muestran los topónimos con sus coordenadas geográficas en un mapa. Este mapa es el que devuelve el sistema a partir del estado final de la pila de desambiguación $\mathbb{S}$. Se encuentra divido en dos partes, en la parte superior se ve la parte correspondiente al *estado de Jalisco*, con sus respectivos topónimos relacionados a este estado. En la parte inferior se encuentra lo relacionado a la *Ciudad de México*. Cada uno de estos topónimos cuenta con dos pines por entidad, el color azul es la opción que el método considera correcta, el color negro indica una ubicación alterna para este topónimo. De acuerdo al ejemplo de la Figura B.1, se identificaron dos topónimos que no estaban en $\mathbb{G}$, por lo que no contaban con coordenadas propias, la forma en la que el método propuesto soluciona esto es asignándolo a la entidad más cercana a ella en la pila $\mathbb{S}$, es por esto

que en la Figura B.2 *Zapatería Juárez* comparte las coordenadas de *Pedregal de Santo Domingo* y *Mueblería Tu Hogar* comparte las coordenadas de Azcapotzalco.

| REGLA | DESCRIPCIÓN | ANTECEDENTE | CONSECUENTE |
|---|---|---|---|
| $R_0$ | El topónimo a ser procesado ($A$) existe en $\mathbb{G}$. $\mathbb{S}$ está vacía. | $P_1 \wedge P_2 \implies$ | $Q_1$ Asignar a $A$ las propiedades geográficas del topónimo encontrado en $\mathbb{G}$. y tiene el nivel jerárquico más alto.<br>$Q_2$ Colocar $A$ en $\mathbb{S}$ |
| $R_1$ | El topónimo a ser procesado ($A$) existe en $\mathbb{G}$. $\mathbb{S}$ no está vacía. $A$ es predecesor del topónimo en el top ($T$) de $\mathbb{S}$. Existe una asociación entre $A$ y $T$. | $P_1 \wedge \neg P_2 \wedge P_3 \wedge P_4 \implies$ | $Q_1$ Asignar a $A$ las propiedades geográficas del topónimo encontrado en $\mathbb{G}$ y tiene el mismo *parent code* que $T$<br>$Q_2$ Colocar $A$ en $\mathbb{S}$ |
| $R_2$ | El topónimo a ser procesado ($A$) existe en $\mathbb{G}$. $\mathbb{S}$ no está vacía. $A$ es predecesor del topónimo en el top ($T$) de $\mathbb{S}$. No existe una asociación entre $A$ y $T$. | $P_1 \wedge \neg P_2 \wedge P_3 \wedge \neg P_4 \implies$ | $Q_1$ Asignar a $A$ las propiedades geográficas del topónimo encontrado en $\mathbb{G}$ y tiene el nivel jerárquico más alto.<br>$Q_2$ Colocar $A$ en $\mathbb{S}$ |
| $R_3$ | El topónimo a ser procesado ($A$) existe en $\mathbb{G}$. $\mathbb{S}$ no está vacía. $A$ no es predecesor del topónimo en el top ($T$) de $\mathbb{S}$. Existe una asociación entre $A$ y $T$. | $P_1 \wedge \neg P_2 \wedge \neg P_3 \wedge P_4 \implies$ | $Q_1$ Asignar $T$ a una nuevo topónimo ($B$)<br>$Q_2$ Eliminar $T$ de $\mathbb{S}$<br>$Q_3$ Asignar a $A$ las propiedades geográficas del topónimo encontrado en $\mathbb{G}$ y tiene el nivel jerárquico más alto.<br>$Q_4$ Colocar $A$ en $\mathbb{S}$<br>$Q_6$ Colocar $B$ en $\mathbb{S}$ |
| $R_4$ | El topónimo a ser procesado ($A$) no existe en $\mathbb{G}$. $\mathbb{S}$ no está vacía. | $\neg P_1 \wedge \neg P_2 \implies$ | $Q_1$ Asignar a $A$ las propiedades geográficas de $T$.<br>$Q_2$ Colocar $A$ en $\mathbb{S}$ |
| $R_5$ | El topónimo a ser procesado ($A$) no existe en $\mathbb{G}$. $\mathbb{S}$ está vacía. | $\neg P_1 \wedge P_2 \implies$ | $Q_1$ Colocar $A$ en $\mathbb{C}$ (Pila de conflictos) |
| $R_6$ | El topónimo a ser procesado ($A$) existe en $\mathbb{G}$. $\mathbb{S}$ no está vacía. $A$ no es predecesor del topónimo en el top ($T$) de $\mathbb{S}$. No existe una asociación entre $A$ y $T$. | $P_1 \wedge \neg P_2 \wedge \neg P_3 \wedge \neg P_4 \implies$ | $Q_1$ Asignar a $A$ las propiedades geográficas del topónimo encontrado en $\mathbb{G}$ y tiene el nivel jerárquico más alto.<br>$Q_2$ Colocar $A$ en $\mathbb{S}$ |
| $R_7$ | No existen más topónimos a procesar, pero $\mathbb{C}$ no está vacía. | $P_5 \wedge \neg P_6 \implies$ | $Q_1$ Colocar todo el contenido de $\mathbb{S}$ en una pila nueva $\mathbb{S}'$, en la que la base de $\mathbb{S}$ sea el top de $\mathbb{S}'$<br>$Q_2$ Extraer el top de $\mathbb{S}'$<br>$Q_3$ Obtener el topónimo con el nivel jerárquico más pequeño asociado al top de $\mathbb{S}'$<br>$Q_4$ Asociar todos los topónimos de $\mathbb{C}$ a ese topónimo. |

Tabla 3.3: Conjunto de reglas empleadas en el método de desambiguación

Llega un cliente a la **Mueblería Tu Hogar**, en la **Ciudad de México** sucursal **Azcapotzalco**. Solicita un comedor que está en oferta, pero pide que se realice un envio a domicilio al poblado de **Ixcatán** en el municipio de **Zapopan**, **Jalisco**.
El problema es que la sucursal encargada de hacer los envíos a esa zona es la de **Pedregal de Santo Domingo**, en el municipio de **Coyoacán**, junto a la **Zapatería Juárez**, pero en esa sucursal el comedor no tiene descuento. La mueblería deberá llegar a un acuerdo con el cliente.

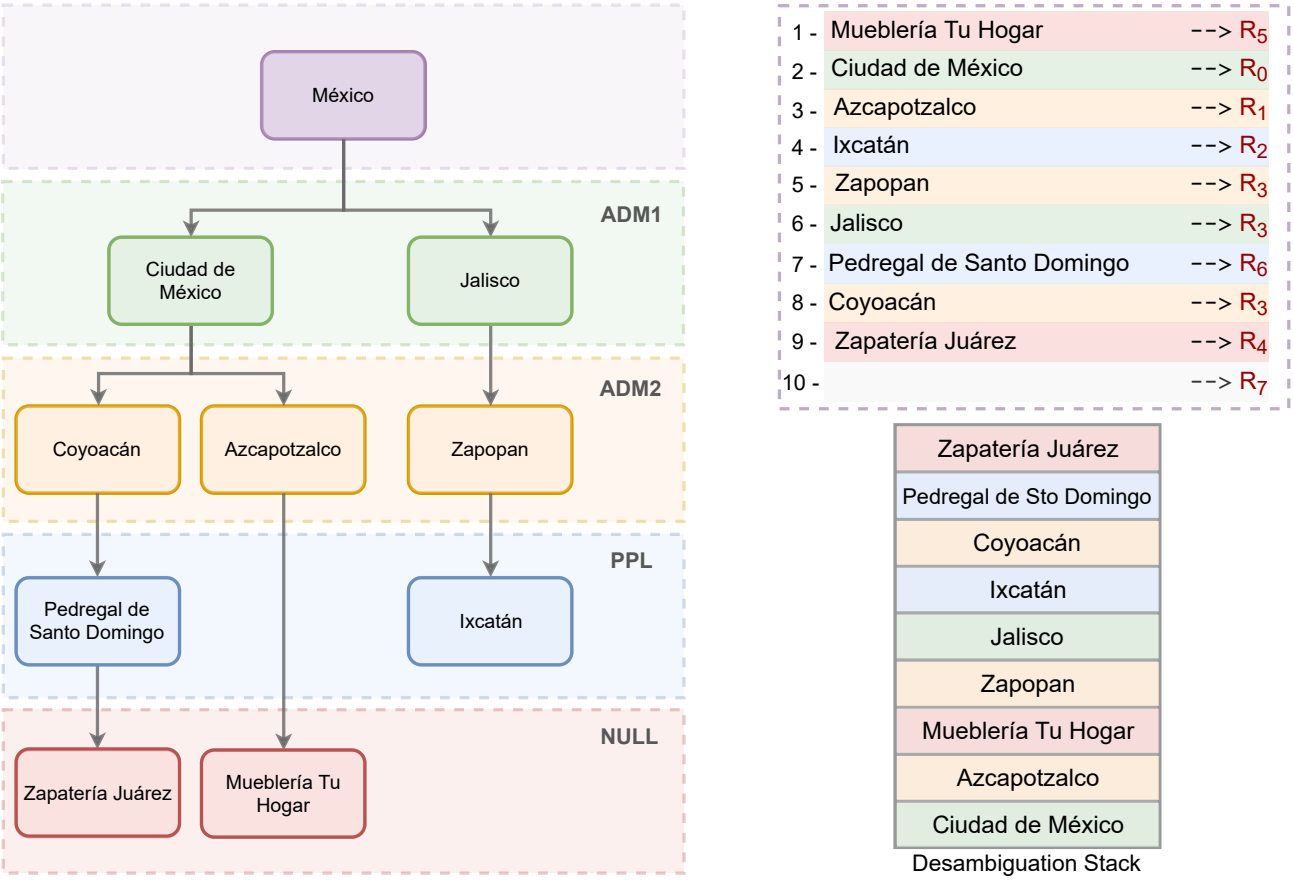


Figura 3.5: Ejemplo de uso del método propuesto

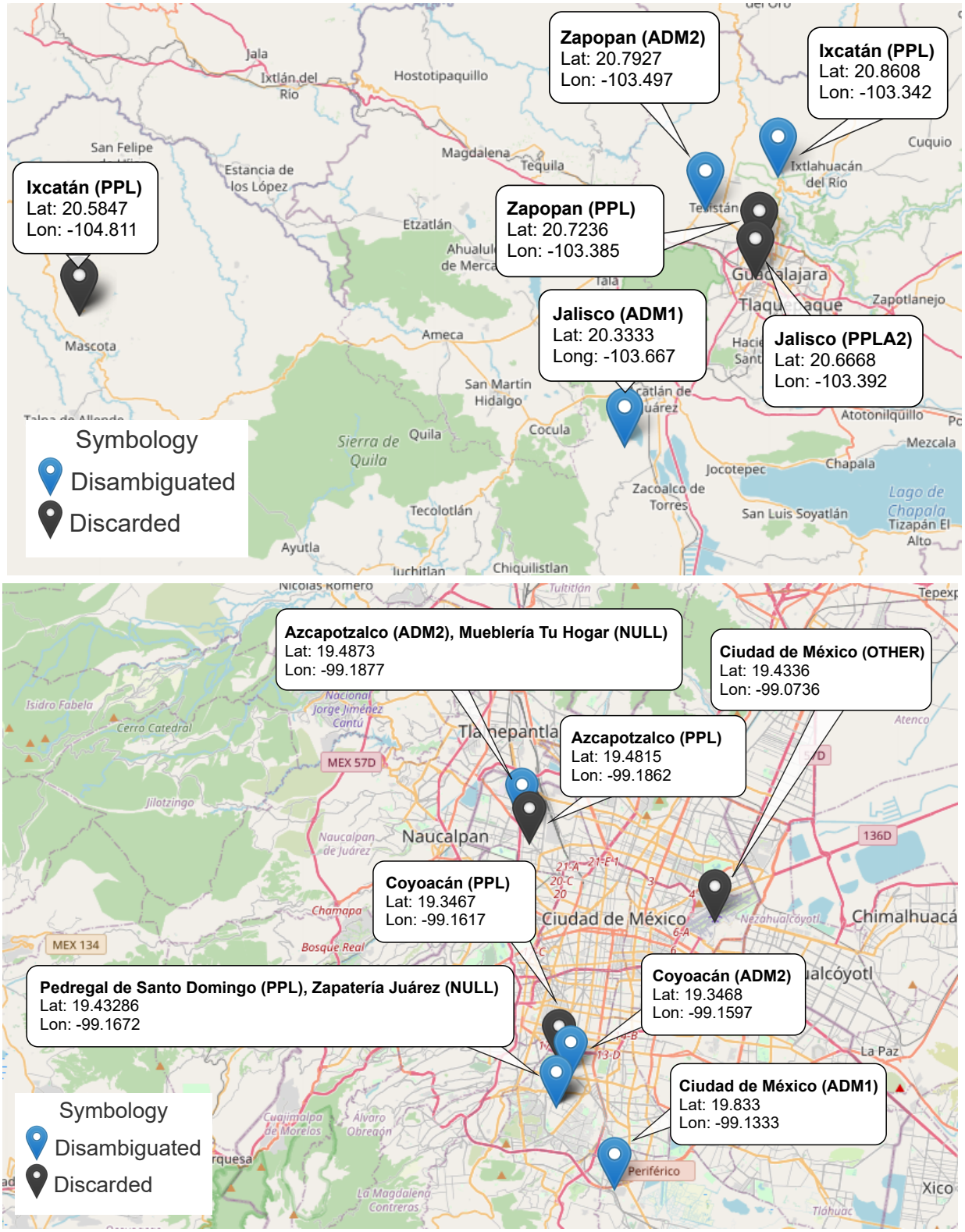


Figura 3.6: Mapa obtenido a partir del ejemplo de la Figura B.1

# 4

# Experimentación y Resultados

En este capítulo se describe la experimentación realizada con el fin de evaluar el desempeño del método propuesto. En primer lugar se incluye una descripción de los materiales utilizados durante la experimentación. Subsecuentemente se describe el diseño experimental y se presentan los resultados obtenidos después del análisis y evaluación de los mismos.

## 4.1 Materiales

A continuación se presenta una descripción de los materiales usados para la etapa de experimentación, se incluyen tanto elementos de software como de hardware, además de elementos complementarios que hacen parte fundamental para esta etapa.

- **Infraestructura:** Ésta se dividió en dos partes:

    1. *Hardware:* Se utilizaron dos equipos, las características de cada uno de ellos se muestran en la Tabla 4.1. El primer equipo utilizado corresponde a un servidor instalado en el

Cinvestav Unidad Tamaulipas, en él se encuentra la aplicación, librerías y gazetteer que permiten desambiguar topónimos en un texto dado. El segundo equipo consiste en una laptop utilizada para la etapa de validación y ajuste del método.

| Equipo | Sistema Operativo | Procesadores | Cores | RAM | Disco |
|---|---|---|---|---|---|
| Servidor | Ubuntu 18 Server | 4 Intel Xeon 2.10 GHz | 48 | 128GB | 1TB HDD |
| Laptop | Windows 10 | 1 Intel Core i5 2.4 GHz | 4 | 6GB | 256GB SSD |

Tabla 4.1: Características de los equipos utilizados para la etapa de experimentación.

2. *Software:* Los elementos de software requeridos para llevar a cabo la etapa de experimentación son los siguientes:
   - Python: Es necesario contar con el intérprete de Python para poder ejecutar las pruebas.
   - MySQL: Para almacenar los resultados de los experimentos.
   - CLIPSPy: Se empleó para conectar Python con CLIPS.
   - PyMySQL: Se empleó esa API para conectar Python con MySQL.

- **Elementos complementarios:** En este apartado se incluyen algunos elementos necesarios para el funcionamiento de la implementación del método propuesto.
  - **GeoNames**[1]**:** Es una base de datos geográfica (gazetteer) que contiene información estructurada acerca de más de 25 millones de topónimos de todo el mundo. Está disponible a través de varios servicios web bajo una licencia Creative Commons. Para este proyecto se hizo uso de los datos relacionados con topónimos de México. GeoNames proporciona diversas características territoriales, entre las cuales se incluyen la latitud, longitud, características del área (feature_class), códigos de características (feature_code), código

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

del estado al que pertenece (admin1_code), código del municipio (admin2_code), población, elevación. Además, en cuanto al topónimo, también proporciona variantes en idiomas, o formas alternas de identificarlo. Todas las coordenadas (latitud y longitud), están en WGS84 (World Geodetic System 1984), que es el Sistema Geodésico Mundial desde 1984.

- **Nominatim**[2]**:** Es un motor de búsqueda de datos de OpenStreetMap[3] (OSM). Utiliza una base de datos PostgreSQL como back-end para almacenar sus datos. Permite hacer dos tipos de búsqueda: 1) búsqueda hacia adelante o geocodificación, es decir, ingresar un nombre o dirección y obtener las coordenadas del mismo y 2) búsqueda inversa o geocodificación inversa, es decir, buscar datos a partir de las coordenadas geográficas. Nominatim está compuesto por tres partes básicas. La primera es importación de datos, que es la encargada de leer los datos de OSM y extraer toda la información útil para la geocodificación. La segunda es la etapa de cálculo o indexación, esta toma los datos y agrega información adicional necesaria para realizar la geocodificación; también permite clasificar los topónimos por importancia, calcular direcciones y el índice de búsqueda. La tercera es la interfaz de búsqueda, esta permite realizar las consultas del usuario, tanto de geocodificación como de geocodificación inversa, busca los datos y regresa al usuario los resultados.

- **GeoparseMX**[4]. Para el etiquetado de dichas noticias se hizo uso de una plataforma de Geoparsing en Español, desarrollada en el CentroGeo Unidad Yucatán[5], llamado GeoparseMX. Esta plataforma permitió ingresar las noticias y asignar manualmente las coordenadas reales de cada uno de los topónimos identificados. Como salida proporciona un archivo JSON que cuenta con el texto etiquetado, el nombre de la entidad, latitud y

[2]https://nominatim.openstreetmap.org
[3]https://www.openstreetmap.org
[4]http://geoparsing.geoint.mx/mx/info
[5]https://www.centrogeo.org.mx

longitud.

## 4.2 Diseño experimental

Para cada corpus la experimentación consistió en dos procesos, que se ilustran en la Figura 4.1:

1. El proceso de experimentación en el que se extraen los topónimos contenidos en cada documento del corpus, los cuales son pasados al proceso de desambiguación (especificado en el Algoritmo 1). Los resultados de este proceso son almacenados en un repositorio de resultados que incluye las coordenadas manualmente asignadas y las coordenadas estimadas de cada topónimo en el documento, así como su categoría geográfica manualmente asignada y su categoria geográfica estimada. Para efectos ilustrativos en la Figura 4.2 se muestran algunas instancias de dicho repositorio.

2. El proceso de evaluación donde, a partir de los resultados experimentales, se calcula un conjunto de métricas de evaluación (Sección 4.2.2) definidas en términos de: 1) aproximación geográfica entre las coordenadas manualmente asignadas y las coordenadas estimadas de los topónimos procesados y 2) precisión en la estimación de la categoría geográfica de los mismos.

Es oportuno recordar que los documentos en cada corpus fueron previamente etiquetados por el módulo reconocedor de entidades geográficas (GNER) para conocer a priori los topónimos contenidos en dichos documentos y sus propiedades geográficas actuales (coordenadas, categoría geográfica).

### 4.2.1 Corpus de prueba

La experimentación se realizó a partir de dos corpus etiquetados.

a) Corpus "El Gráfico", está compuesto por 300 noticias extraídas del periódico "El Gráfico"[6], dichas noticias están relacionadas con temas de *secuestro y enfrentamientos*. En estos

[6]https://www.elgrafico.mx

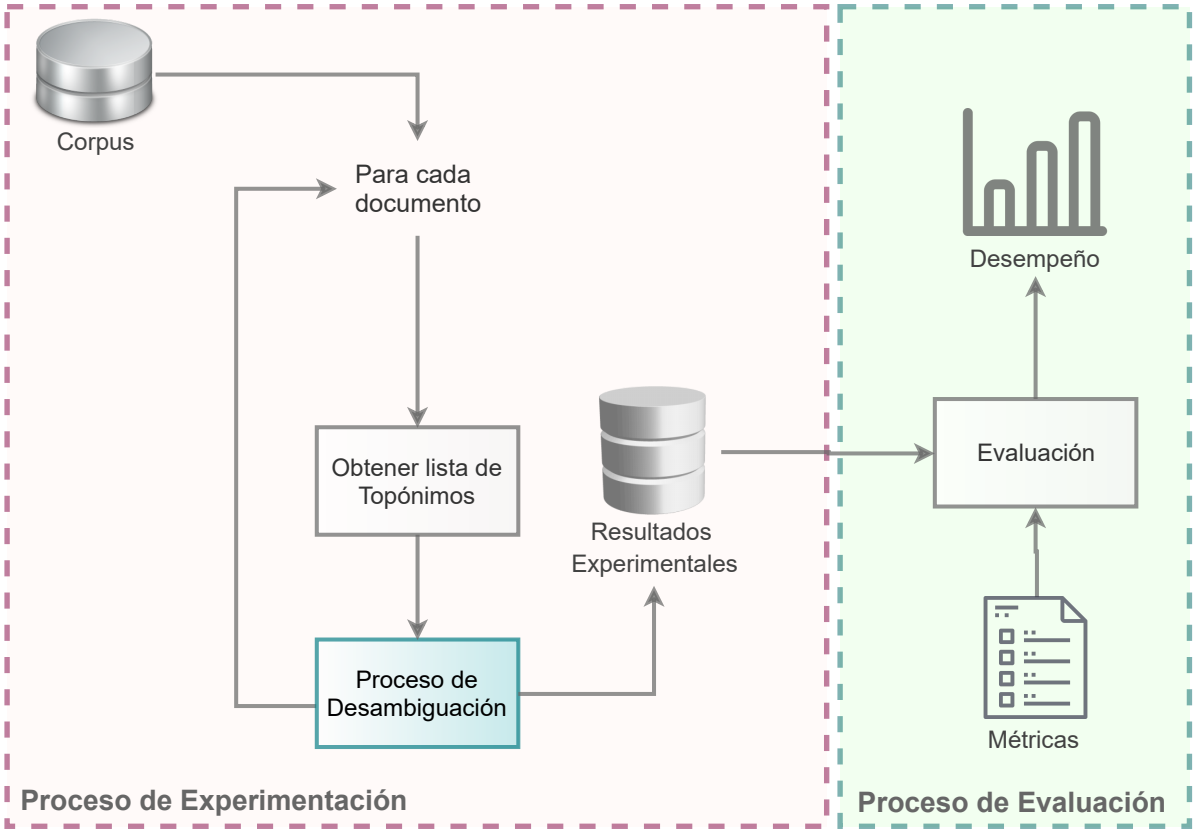


Figura 4.1: Diseño Experimental

documentos se identificó un total de 1956 topónimos.

b) Corpus "Newspapers", está compuesto por 350 noticias de los periódicos "La Jornada"[7] y "El Universal"[8], tales noticias también están relacionadas con temas de *secuestros y enfrentamientos*, dentro de estas se identificaron 2288 topónimos.

### 4.2.2 Métricas

A continuación se describen las métricas usadas durante el proceso de evaluación. La primera es la *aproximación geográfica*, definida como la distancia entre las coordenadas estimadas por el método y las coordenadas manualmente asignadas. La segunda es la *precisión categórica*, que representa la

[7] https://www.jornada.com.mx
[8] https://www.eluniversal.com.mx

| Topónimo | Coordenadas manualmente asignadas | Coordenadas Estimadas | Característica Geográfica manualmente asignada | Característica Geográfica Estimada |
|---|---|---|---|---|
| Cuernavaca | 18.9331, -99.2599 | 18.9261, -99.2308 | PPLA | PPLA |
| Naucalpan | 19.4785, -99.2396 | 19.4785, -99.2396 | PPLA2 | PPLA2 |
| Estado de México | 19.4839, -99.6899 | 19.3667, -99.6667 | ADM1 | ADM1 |

Figura 4.2: Ejemplo de tabla con resultados experimentales

efectividad en la asignación de una categoría geográfica basada en orden administrativo (estados, ciudades, municipios, etc.).

- **Aproximación geográfica:**

  Esta medida permite conocer qué tan cerca se encuentran las coordenadas estimadas por el método propuesto con respecto a las coordenadas manualmente asignadas. Para ello se utilizó la distancia *Haversine*, la cual permite conocer la distancia entre dos puntos $p_1$ y $p_2$ en una esfera, definidos en términos de latitud y longitud (ilustrada en la Figura 4.3). Esta distancia se define como:

$$\text{Haversine} = \cos^{-1}(\text{sen}(\phi_1) * \text{sen}(\phi_2) + \cos(\phi_1) * \cos(\phi_2) * \cos(\lambda_1) - \lambda_2) * r \quad (5)$$

  donde $\phi_i$ y $\lambda_i$ corresponden a la latitud y longitud del punto $p_i$ y $r$ se refiere al radio de la tierra (6371 km)

- **Precisión categórica:** Esta métrica se define con base en una *matriz de confusión* (Tabla 4.2) que se obtiene a partir de los resultados de la experimentación. Por cada topónimo desambiguado se determina si el método acertó o no; esto se indica como *correcto* o *incorrecto*. En dicha matriz, las clases corresponden a 8 categorías geográficas definidas en la Tabla 3.2.

Figura 4.3: Distancia Haversine

A partir de los valores correctos e incorrectos de la desambiguación, por cada topónimo se obtienen los valores correspondientes a verdaderos positivos (VP), verdaderos negativos (VN), falsos positivos (FP) y falsos negativos (FN); los cuales se representan en la matriz de confusión. A partir de estos valores es posible definir las siguientes métricas:

| | | Predicción | |
|---|---|---|---|
| | | Positivo | Negativo |
| Real | Positivo | True Positive (TP) | True Negative (TN) |
| | Negativo | False Positive (FP) | False Negative (FN) |

Tabla 4.2: Matriz de confusión para dos clases

- **Accuracy** (Exactitud), se refiere al número de topónimos asignados correctamente con respecto a todas las asignaciones, tanto correctas como incorrectas.

$$\text{Accuracy} = \frac{VP+VN}{VP+VN+FP+FN} \qquad (1)$$

- **Precision** (Precisión), se refiere al número de topónimos asignados correctamente sobre el número total de topónimos.

$$\text{Precision} = \frac{VP}{VP+FP} \qquad (2)$$

- **Recall** (Cobertura), se refiere al número de topónimos asignados correctamente sobre el número total de topónimos que deberían haberse devuelto.

$$\text{Recall} = \frac{VP}{VP+FN} \qquad (3)$$

- **F-measure** (Medida-F), es la media armónica entre la precisión y la cobertura. Toma en cuenta la cantidad de resultados positivos y negativos devueltos. Sus valores se encuentran entre 0 y 1.

$$\text{F-measure} = \frac{2VP}{2VP+FP+FN} \qquad (4)$$

Dado que en la desambiguación se pueden tener hasta 8 categorías, es necesario promediar los diferentes resultados obtenidos para conseguir un solo valor que represente el método; para ello existen algunas formas, de las cuales solo utilizaremos dos, macro y micro. Macro considera un promedio sobre el total de clases denotado como $n$, a diferencia de micro que toma en cuenta los resultados obtenidos por clase. A continuación se presentan las fórmulas para calcular los micro y macro valores para cada una de las métricas anteriores.

- $\text{Micro-Precision} = \frac{\sum i^n TP_i}{\sum i^n TP_i + \sum i^n FP_i}$ (5)
- $\text{Macro-Precision} = \frac{\sum i^n Precision_i}{n}$ (6)
- $\text{Micro-Recall} = \frac{\sum i^n TP_i}{\sum i^n TP_i + \sum i^n FN_i}$ (7)
- $\text{Macro-Recall} = \frac{\sum i^n Recall_i}{n}$ (8)
- $\text{Micro-F-measure} = 2 * \frac{Micro-Precision * Micro-Recall}{Micro-Precision + Micro-Recall}$ (9)
- $\text{Macro-F-measure} = 2 * \frac{Macro-Precision * Macro-Recall}{Macro-Precision + Macro-Recall}$ (10)

### 4.2.3 Nominatim como punto de referencia

Para complementar la evaluación de desempeño del método propuesto relativo a la proximidad geográfica se utilizó el motor de búsqueda *Nominatim*. El propósito fue obtener, a través de esta herramienta, una estimación de las coordenadas geográficas de los topónimos contenidos en el corpus. A partir de dicha estimación se calculó la proximidad geográfica. Esta proximidad se comparó con la

obtenida por el método propuesto. Dicha comparación puede ser interpretada como una medida de efectividad del método respecto a una herramienta de georreferenciación ampliamente conocida.

## 4.3 Análisis de resultados

Una vez ejecutadas las pruebas se obtuvieron los resultados mostrados en las gráficas y tablas siguientes, mediante los cuales es posible conocer el desempeño del método propuesto. Estos resultados se dividen en los dos corpus empleados.

### 4.3.1 Corpus "El Gráfico"

Este corpus cuenta con un total de 1956 topónimos, de los cuales el método propuesto fue capaz de georreferenciar el $100\%$, a diferencia de Nominatim que sólo asigna georreferencias al $69{,}83\%$ del total de topónimos.

A continuación se describen de manera detallada los resultados obtenidos a partir de las métricas definidas anteriormente.

- **Aproximación geográfica**

  Como ya se mencionó anteriormente, cada topónimo cuenta con un par de coordenadas esperadas, éstas son contrastadas con las coordenadas estimadas por los dos métodos: el método propuesto y Nominatim.

  La distancia en kilómetros obtenida entre la coordenada esperada y la estimada se muestra en el histograma presentado en la Figura 4.4, el método propuesto se representa con líneas inclinadas, Nominatim con líneas horizontales. Esta distancia representa qué tan cerca estuvo cada método del topónimo que está georreferenciando. En el eje horizontal se muestran los rangos en kilómetros para cada topónimo. En el eje vertical la cantidad de topónimos reconocidos por cada uno de los métodos. Los resultados de esta gráfica se calcularon con respecto a 1956

topónimos (total de topónimos en el corpus) de los cuales el método propuesto georreferencia 1100 topónimos en un rango de 0 a 5 kilómetros, lo cual representa un $56\%$ del total, a diferencia de Nominatim que georreferencia 446 topónimos en el mismo rango, lo que equivale al $22{,}80\%$.

Por otro lado, en cuanto a los topónimos con una georreferencia mayor a los 100 kilómetros, el método propuesto georreferenció a 278 topónimos ($14{,}21\%$) y Nominatim 446 topónimos ($22{,}80\%$). En el rango "No georreferenciado" se encuentra el total de topónimos a los que no se les asigna ninguna coordenada. En éste, el método propuesto tiene 0 debido a que es capaz de asignar una georreferencia cada topónimo recibido, en contraste con Nominatim, que no asigna coordenadas a 586 topónimos.

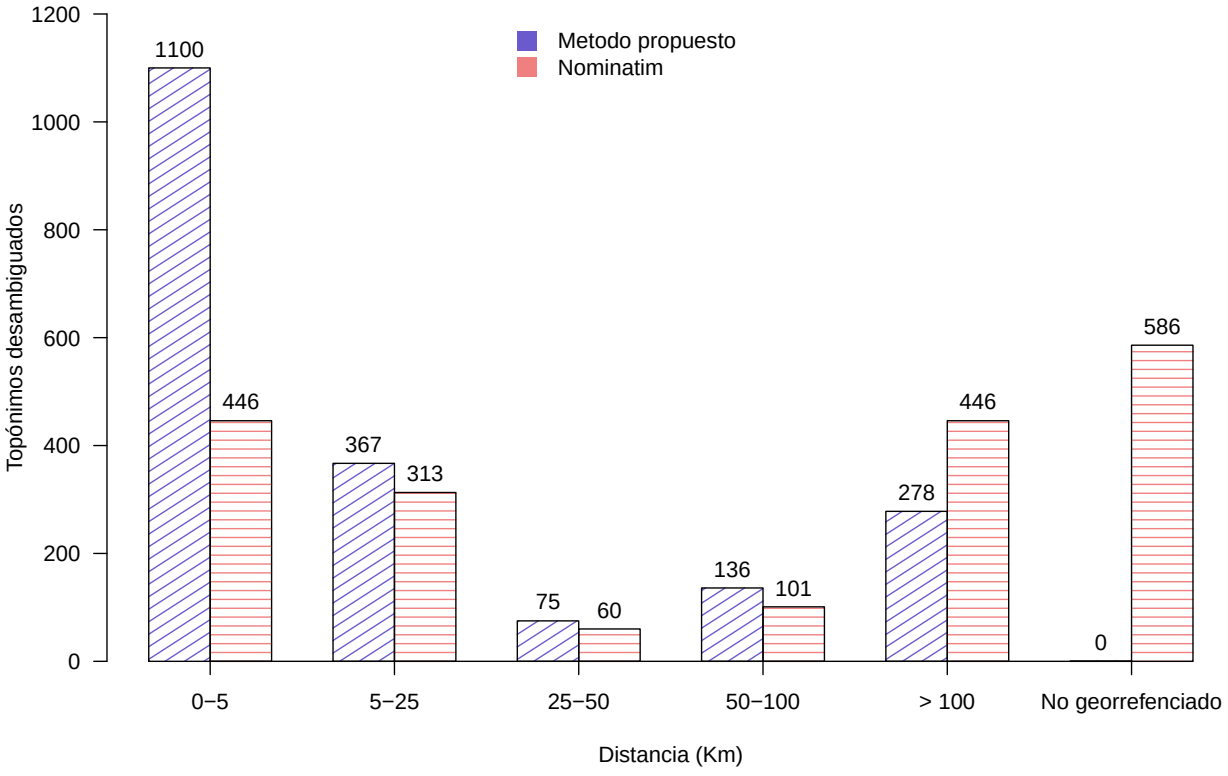


Figura 4.4: Resultados obtenidos con respecto al total de topónimos en el corpus "El Gráfico".

A diferencia de la gráfica descrita con anterioridad, en la Figura 4.5 se muestra la frecuencia de aparición de los topónimos en cada uno de los rangos definidos con respecto al total de topónimos identificados por cada método. Es decir, los porcentajes presentados en la gráfica de la izquierda (Método Propuesto) están calculados con respecto a 1956 topónimos, debido a que es el total de topónimos georreferenciados. Los porcentajes presentados en la gráfica de la derecha (Nominatim) están calculados con respecto a 1366 topónimos, que fue el total que este método pudo georreferenciar. En esta gráfica se observa que el método propuesto concentra la mayor parte de los topónimos en el primer rango (0 a 5 Km) con un $56{,}23\%$, a diferencia de Nominatim con $32{,}65\%$ de topónimos en este rango. En el caso de Nominatim el $32{,}74\%$ del total se encuentran en un rango de más de 100 Km con respecto al resultado esperado, para este mismo rango el método propuesto coincide con un $14{,}2\%$ de topónimos.

En estas dos gráficas se puede observar que el método propuesto obtuvo un mejor desempeño con el corpus "El Gráfico" en comparación con Nominatim. Esto se concluye debido a que la mayoría de topónimos georreferenciados por el método propuesto caen en el rango más cercano. Además, como ya se mencionó, el método propuesto fue capaz de asignar georreferencias a todas las entidades presentadas aunque estas no se encuentren en el gazetteer empleado.

En la Tabla 4.3 se muestra un resumen del número de topónimos y el porcentaje con respecto a la cantidad de topónimos del corpus para cada uno de los rangos definidos. Además de un resumen de los resultados de cada método con relación a la cantidad de topónimos georreferenciados por cada uno de ellos. Se puede observar que para cada uno de estos rangos el desempeño del método propuesto queda por encima de Nominatim. Además, al tomar en cuenta solo los topónimos que puede georreferenciar cada método, el desempeño de Nominatim mejora pero no alcanza los resultados del método propuesto.

- **Precisión categórica**

Las métricas de precisión categórica están basadas en los valores de la *matriz de confusión*

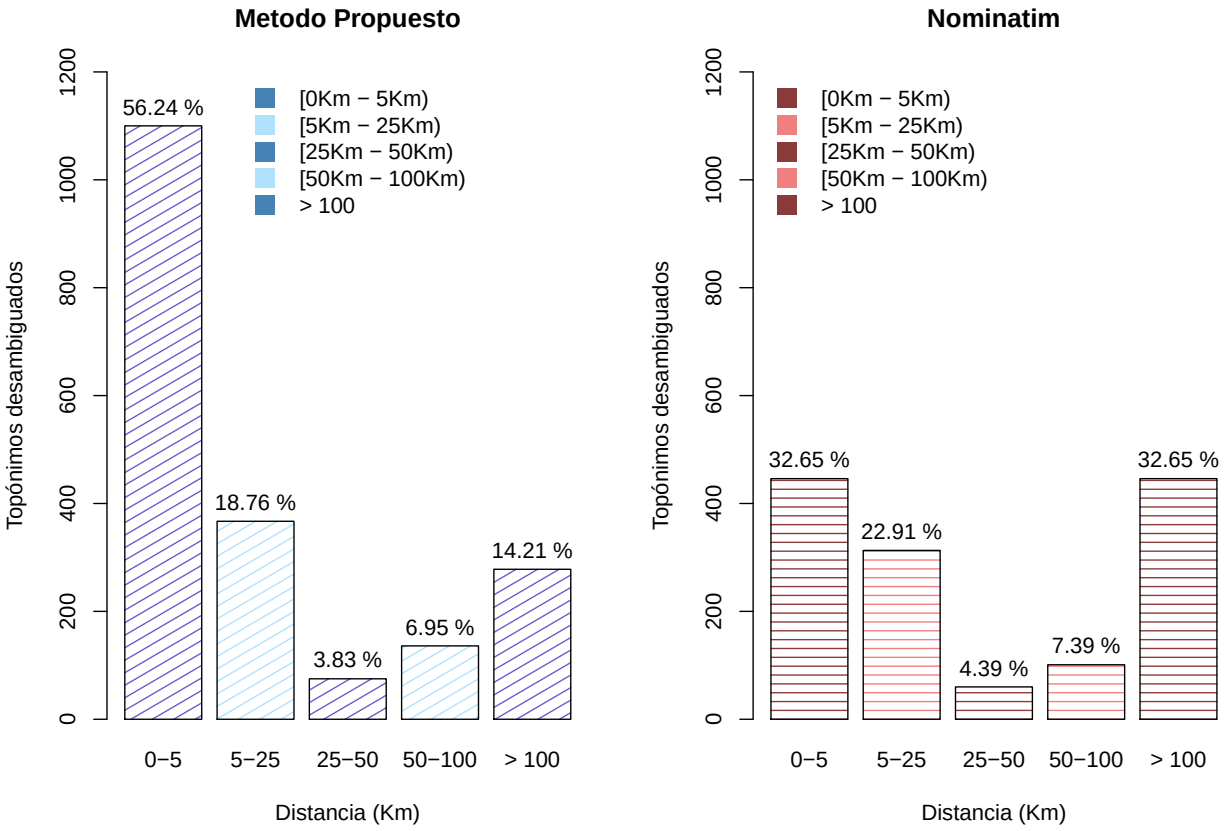


Figura 4.5: Resultados obtenidos con respecto al total de topónimos identificados por cada método del corpus "El Gráfico".

obtenida en el proceso de experimentación, donde las filas representan las instancias estimadas por cada categoría geográfica, mientras que las columnas representan las instancias en la categoría geográfica actual. En la Tabla 4.4 se muestra la matriz de confusión para el corpus "El Gráfico", a partir de la cual se calculan las métricas Accuracy, Recall, Precision y F-measure de la Tabla 4.5.

Se puede observar un valor de *accuracy* de 0.9089 que indica que el $90\%$ de los topónimos presentados han sido asignados correctamente. Para *micro-precision* y *macro-precision* se tiene un valor de 0.9089 y 0.7634 respectivamente, lo que indica que el $90\%$ de las veces que el método asignó un topónimo a una categoría, este topónimo realmente correspondía a ella;

| | Con respecto al total de topónimos | | | | Con respecto a los topónimos georreferenciados por método | | | |
|---|---|---|---|---|---|---|---|---|
| | Metodo Propuesto | | Nominatim | | Metodo Propuesto | | Nominatim | |
| Rangos en Km | N. Topónimos | Porcentaje | N. Topónimos | Porcentaje | N. Topónimos | Porcentaje | N. Topónimos | Porcentaje |
| [0 - 5) | 1100 | 56.24 % | 446 | 22.80 % | 1100 | 56.24 % | 446 | 32.65 % |
| [5 - 25) | 367 | 18.76 % | 313 | 16 % | 367 | 18.76 % | 313 | 22.91 % |
| [25 - 50) | 75 | 3.83 % | 60 | 3.06 % | 75 | 3.83 % | 60 | 4.39 % |
| [50 - 100) | 136 | 6.95 % | 101 | 5.16 % | 136 | 6.95 % | 101 | 7.39 % |
| [100 - 2000) | 278 | 14.21 % | 446 | 22.80 % | 278 | 14.21 % | 446 | 32.65 % |
| Sin georreferencia | 0 | 0 % | 594 | 30.36 % | - | - | - | - |

Tabla 4.3: Resumen de los resultados obtenidos del corpus "El Gráfico", con respecto al total de topónimos y a los topónimos georreferenciados por método.

| | | *Actual* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Clase | **ADM1** | **PPLA** | **ADM2** | **PPLA2** | **PPL** | **LCTY** | **OTHER** | **NULL** |
| *E* | **ADM1** | 222 | 1 | 2 | 0 | 8 | 0 | 1 | 2 |
| *s* | **PPLA** | 3 | 49 | 2 | 0 | 0 | 0 | 1 | 0 |
| *t* | **ADM2** | 6 | 5 | 655 | 0 | 16 | 0 | 1 | 3 |
| *i* | **PPLA2** | 2 | 3 | 0 | 73 | 0 | 0 | 0 | 1 |
| *m* | **PPL** | 2 | 0 | 6 | 1 | 265 | 0 | 1 | 11 |
| *a* | **LCTY** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| *d* | **OTHER** | 14 | 0 | 4 | 1 | 4 | 0 | 26 | 5 |
| *a* | **NULL** | 17 | 5 | 17 | 2 | 28 | 0 | 2 | 488 |

Tabla 4.4: Matriz de confusión del corpus "El Gráfico".

además que con respecto a las otras categorías la asignación correcta fue del $76\%$. En el caso de *micro-recall* y *macro-recall* en los que el valor es 0.9089 y 0.7489 respectivamente, indica la proporción de veces en las que el método propuesto asignó un topónimo a una categoría correcta con respecto al total de categorías esperadas, tanto para la misma categoría como para las demás. Finalmente *micro-F-measure* y *macro-F-measure* es una combinación de los resultados entre *precision* y *recall* micro y macro, esto indica que el $90\%$ de las veces el método propuesto realizó las asignaciones que se esperaban para una misma categoría y el $70\%$ de acierto con respecto a las otras categorías. En general, con base en estos resultados, se puede ver que el método propuesto presenta un buen desempeño en la asignación de georreferencias a topónimos del corpus "El Gráfico".

| | Accuracy | Precision | | Recall | | F-measure | |
|---|---|---|---|---|---|---|---|
| | | Micro | Macro | Micro | Macro | Micro | Macro |
| Corpus "El Gráfico" | 0.9089 | 0.9089 | 0.7634 | 0.9089 | 0.7489 | 0.9089 | 0.7493 |

Tabla 4.5: Resultados de la evaluación por precisión categórica del corpus "El Gráfico"

### 4.3.2 Corpus "Newspapers"

Este corpus cuenta con un total de 2288 topónimos dividido en 350 noticias. En este caso, al igual que en el corpus anterior, el método propuesto fue capaz de georreferenciar el $100\%$ de los topónimos que se le presentan, a diferencia de Nominatim, que en esta ocasión sólo asigna georreferencias al $74{,}78\%$.

- **Aproximación geográfica**

  Las diferencias entre las coordenadas reales y las coordenadas estimadas tanto por el método propuesto como por Nominatim se presentan en la Figura 4.6. Aquí los resultados del método propuesto se muestran con líneas inclinadas y los de Nominatim con líneas horizontales. En el eje horizontal están colocados los rangos en kilómetros en los que cada método estuvo cerca de georreferenciar a un topónimo, y en el eje vertical la frecuencia de aparición de los topónimos en cada uno de los rangos. Como se puede observar en el primer rango, que es el de 0 a 5 kilómetros, el método propuesto georreferencia 1815 topónimos que corresponden al $79{,}33\%$, a diferencia de Nominatim que georreferencia 932 topónimos que corresponden al $40{,}73\%$. En el rango de más de 100 kilómetros el método propuesto georreferencia 264 topónimos y Nominatim 95, a simple vista podríamos deducir que el método propuesto colocó erróneamente estos topónimos, pero al hacer un análisis a fondo se puede notar que la mayoría de éstos son los que no se encuentran en el gazetteer, por lo que el método propuesto no las georreferencia correctamente. En el último rango, el de no georreferenciado, el método propuesto no tiene ningún topónimo, mientras que Nominatim 577.

  En la Figura 4.7 del lado izquierdo se muestran los resultados obtenidos por el método propuesto

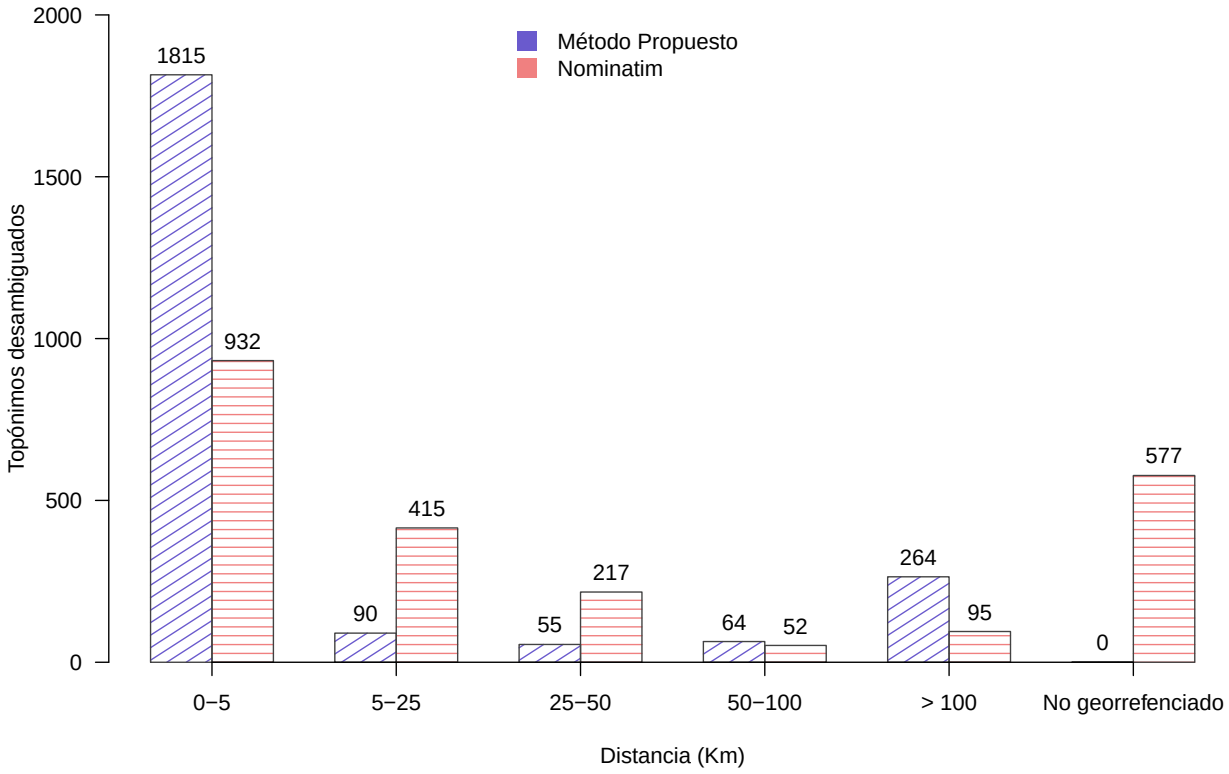


Figura 4.6: Resultados obtenidos con respecto al total de topónimos en el corpus "Newspapers".

y del lado derecho los resultados obtenidos por Nominatim. En el eje horizontal de las dos gráficas se encuentran los rangos en los que se dividieron las distancias obtenidas hacia los topónimos y en el eje vertical se encuentra el total de topónimos que corresponden en cada uno de estos rangos. Estos porcentajes están calculados a partir del total de topónimos en el corpus "Newspaper" que pudo georreferenciar cada método; en el caso del método propuesto son 2288 y en el caso de Nominatim 1711.

Se puede observar que el método propuesto georreferencia el $79{,}33\%$ de los topónimos en el rango de 0 a 5 kilómetros. En los otros rangos se puede ver que los porcentajes son menores al $5\%$, a excepción del de más de 100, que alcanza un $11{,}54\%$. Nominatim georreferencia el $54{,}47\%$ en el primer rango, y las demás categorías presentan un mayor porcentaje en

comparación con las obtenidas por el método propuesto. Aunque en el caso del rango más de 100 tiene $5{,}55\%$, que es menos que los resultados obtenidos por el método propuesto. En general, los resultados del método propuesto se mantienen por encima de los obtenidos por Nominatim.

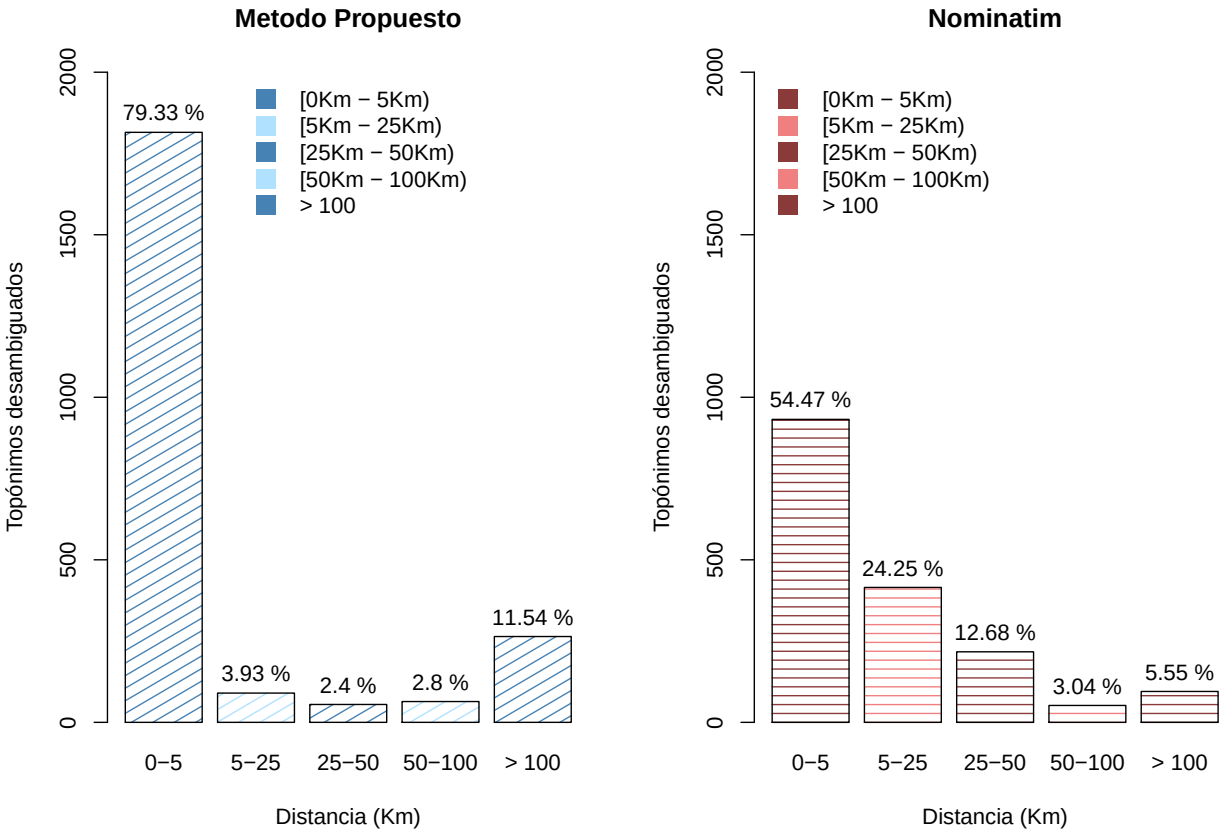


Figura 4.7: Resultados obtenidos con respecto al total de topónimos identificados en el corpus "Newspapers" por cada método.

En la Tabla 4.6 se muestra un resumen de las gráficas anteriores, se calcula un porcentaje a partir del total de topónimos para cada rango. A su vez, estos porcentajes son calculados tanto para la cantidad de topónimos del corpus en general, como para cada uno de los rangos definidos.

- **Precisión categórica**

| | Con respecto al total de topónimos | | | | Con respecto a los topónimos georreferenciados por método | | | |
|---|---|---|---|---|---|---|---|---|
| | Método Propuesto | | Nominatim | | Método Propuesto | | Nominatim | |
| Rangos en Km | N. Topónimos | Porcentaje | N. Topónimos | Porcentaje | N. Topónimos | Porcentaje | N. Topónimos | Porcentaje |
| [0 - 5) | 1815 | 79.33 % | 932 | 40.73 % | 1815 | 79.33 % | 932 | 53.47 % |
| [5 - 25) | 90 | 3.93 % | 415 | 18.14 % | 90 | 3.93 % | 415 | 24.25 % |
| [25 - 50) | 55 | 2.4 % | 217 | 9.48 % | 55 | 2.4 % | 217 | 12.68 % |
| [50 - 100) | 64 | 2.8 % | 52 | 2.27 % | 64 | 2.8 % | 52 | 3.04 % |
| [100 - 2000) | 264 | 11.54 % | 95 | 4.15 % | 264 | 11.54 % | 95 | 5.55 % |
| Sin georreferencia | 0 | 0 % | 577 | 25.22 % | - | - | - | - |

Tabla 4.6: Resumen de los resultados obtenidos en el corpus "Newspapers" con respecto al total de topónimos y a los topónimos georreferenciados por cada método.

Al igual que en el caso del corpus anterior, se obtiene una matriz de confusión de la experimentación con el método propuesto, en la Tabla 4.7 se representa dicha matriz. Las columnas representan instancias esperadas, mientras que las filas instancias estimadas por categoría geográfica. Las etiquetas tanto de las columnas como de las filas se pueden consultar en la Tabla 3.2.

| | | *Actual* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Clase | **ADM1** | **PPLA** | **ADM2** | **PPLA2** | **PPL** | **LCTY** | **OTHER** | **NULL** |
| *E* | **ADM1** | 823 | 5 | 9 | 0 | 3 | 0 | 1 | 3 |
| *s* | **PPLA** | 7 | 102 | 0 | 0 | 0 | 0 | 0 | 0 |
| *t* | **ADM2** | 31 | 0 | 691 | 5 | 13 | 0 | 4 | 6 |
| *i* | **PPLA2** | 0 | 0 | 4 | 42 | 0 | 0 | 0 | 0 |
| *m* | **PPL** | 19 | 5 | 13 | 4 | 305 | 0 | 0 | 21 |
| *a* | **LCTY** | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *d* | **OTHER** | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| *a* | **NULL** | 14 | 6 | 24 | 3 | 13 | 0 | 0 | 112 |

Tabla 4.7: Matriz de confusión corpus "Newspapers"

En la Tabla 4.8 se pueden ver los valores calculados para las métricas Accuracy, Recall, Precision y F-measure a partir de la matriz de confusión. Se puede ver que los resultados obtenidos para cada métrica en el caso de una misma categoría se mantiene en un $90\%$, al igual que en el corpus anterior. Para las métricas *macro* se tienen valores distintos, pero en general se mantiene un buen desempeño similar al corpus anterior. Se puede concluir que el $76\%$ de las asignaciones se realizó de manera correcta con respecto a otras categorías.

En cuanto a este segundo corpus el método propuesto conserva, en todas las métricas, el buen desempeño mostrado al evaluar el corpus anterior.

| | Accuracy | Recall | | Precision | | F-measure | |
|---|---|---|---|---|---|---|---|
| | | Micro | Macro | Micro | Macro | Micro | Macro |
| Corpus 2 | 0.9069 | 0.9069 | 0.8896 | 0.9069 | 0.7662 | 0.9069 | 0.7831 |

Tabla 4.8: Resultados de la evaluación por precisión categórica del corpus "Newspapers"

# 5

# Conclusiones

En este capítulo se presenta un resumen del trabajo realizado que sintetiza el método propuesto y resultados obtenidos; las contribuciones obtenidas después de evaluar una implementación del método propuesto; las limitantes del método propuesto en el contexto del trabajo, la tarea desarrollada y los resultados obtenidos. Asimismo se describe el trabajo futuro que consideramos que podría abordarse a corto y mediano plazo con el trabajo desarrollado.

## 5.1 Resumen

Actualmente la desambiguación de topónimos es una situación presente en muchos ámbitos de la vida cotidiana. Esta situación se presenta cuando dos o más topónimos comparten un mismo nombre pero cada uno debe tener sus propias coordenadas geográficas. El problema se vuelve mucho más complejo si no se tienen los recursos lingüísticos que den soporte a la desambiguación conteniendo información de topónimos muy puntuales de una región.

Este problema de geoparsing consta de dos partes. La primera parte es el reconocimiento de

topónimos y la segunda la desambiguación de los mismos. En general, el problema consiste en asociar topónimos con sus respectivas coordenadas.

En este trabajo de investigación se propuso un método de desambiguación de topónimos basado en la manera en que los humanos resuelven este problema. El método toma como entrada texto en Español con los topónimos identificados, que es la salida de un reconocedor de entidades geográficas y asigna, sin ambigüedad, las coordenadas geográficas de los topónimos del texto. El método está compuesto de las siguientes etapas: generador de hechos, motor de inferencia y ejecutor de acciones.

A partir de la definición del método se realizó la implementación de un prototipo funcional para evaluar la viabilidad del método propuesto. Este prototipo se implementó con Python, datos de OpenStreetMap y un gazetteer de topónimos de México. Se realizó la evaluación del método empleando dos corpus que incluyen topónimos específicos de México. Con la experimentación realizada se corroboró la hipótesis planteada inicialmente: *Es posible asignar propiedades geográficas a topónimos presentes en un texto en español mediante un método de asociación de grano fino, tomando en cuenta el contexto del texto.*

Con lo anterior se demostró que el método es viable para desambiguar topónimos en textos en Español. Los resultados obtenidos son buenos y muy prometedores para seguir trabajando sobre el método. Esto abre las puertas a más pruebas y desarrollos con la idea inicial del método.

## 5.2 Contribuciones

El método de desambiguación propuesto permite asignar coordenadas geográficas a cada topónimo mencionado en un texto en Español, eliminando posibles ambigüedades entre ellos. Esto lo realiza mediante un enfoque basado en reglas que permite emular la forma en que las personas infieren la localización de los topónimos en un texto, con base en el contexto. Las cualidades del método propuesto son:

a) Cuando los topónimos no se encuentran en el gazetteer, el método propuesto es capaz de

asignar las coordenadas más probables. Esto permite enriquecer el gazetteer.

b) A partir de la experimentación realizada se puede ver que el método propuesto permite obtener coordenadas geográficas de topónimos en rangos cercanos a lo esperado.

c) Se comprobó que el método supera en efectividad a Nominatim, que es una de las herramientas más usadas para conocer coordenadas de topónimos de OpenStreetMap.

Las contribuciones de este trabajo son las siguientes:

1. La definición de un método de desambiguación de topónimos presentes en textos en Español.

2. Un prototipo funcional que implementa el algoritmo del método propuesto.

Desde un punto de vista práctico, el método propuesto puede ser de utilidad en aplicaciones en las cuales se requiera extraer información geográfica a partir de texto. Por ejemplo, en redes sociales, correos electrónicos, notas de prensa o documentos jurídicos es de gran importancia identificar topónimos e inferir su ubicación geográfica como apoyo a la toma de decisiones. Incluso, el método propuesto podría ser de utilidad, en sistemas de emergencia, sistemas de asistencia vial o asistencia médica, en los cuales las llamadas recibidas podrían ser transformadas a texto, a partir del cual sería posible identificar topónimos y asignarles las coordenadas geográficas más probables, con el propósito de inferir información geográfica que pudiera ser de ayuda a la atención de usuarios, más allá de la ubicación inferida a partir del dispositivo telefónico desde el cual se genera la llamada.

## 5.3 Limitantes

Si bien el método propuesto e implementado dio muy buenos resultados en la experimentación, superando incluso las expectativas iniciales, algunos aspectos teóricos y prácticos no se pudieron abordar por debido a restricciones de tiempo. Las limitantes más destacables se listan a continuación:

1. Aumentar el número de corpus de pruebas para realizar más experimentación. El método propuesto se probó con dos corpus dando muy buenos resultados. Sería deseable reafirmar la efectividad del método propuesto probando con corpus de dominios diferentes al de noticias. Esto representa un desafío grande por la poca disponibilidad de corpus etiquetados a grano fino (sobre topónimos muy específicos).

2. Aunque el diseño y la implementación del método es agnóstica al lenguaje, la evaluación del método solo incluyó corpus en Español. Por lo que una evaluación con corpus en otros lenguajes sería de mucha utilidad para generalizar la efectividad del método propuesto. Para extenderlo a otro lenguaje es preciso entrenar nuevamente el reconocedor de entidades geográficas con documentos con el nuevo lenguaje y ampliar el gazetteer a otros países.

## 5.4 Trabajo futuro

Los resultados de la experimentación realizada con el prototipo que implementa el método propuesto corroboraron la hipótesis planteada inicialmente. No obstante, a partir de las limitantes del trabajo descritas en la subsección anterior, consideramos que el método podría mejorarse a corto y mediano plazo con la recopilación y etiquetado de documentos (corpus) distintos a los utilizados en la experimentación:

a) Textos en Español que incluyan otros dominios de interés. Es decir, que no sean solo noticias si no cualquier otro texto que contenga topónimos

b) Textos en Español de países distintos a México. Esto implicaría:

   I- Entrenar el reconocedor de entidades nombradas con los textos del país de interés

   II- Emplear un gazetteer que contenga información de los topónimos de dicho país

   III- Contar con información de la división administrativo-geográfica del país en cuestión

c) Similar al inciso anterior pero con textos en idiomas distintos al Español.

d) Liberar la herramienta implementada para su uso público. Actualmente la herramienta sólo es accesible de manera local.

e) Acoplar el módulo de desambiguación al módulo reconocedor de entidades geográficas existente (GeoparseMx).

# A

# Manual de ClipsPy

CLIPS es un lenguaje de programación basado en reglas, desarrollado en C, inicialmente en el Centro Espacial Johnson de la NASA. Proporciona funciones de programación procesal y orientada a objetos [44]. Es utilizado comúnmente para Inteligencia Artificial ya que proporciona un entorno de desarrollo para la ejecución de sistemas expertos. Un sistema experto es un software que emula el comportamiento humano en cuanto a la solución de un problema. Cuenta con tres elementos importantes:

- *Hechos.* Se refiere a una lista de elementos que describen el estado actual del sistema, una serie de hechos es capaz de activar una regla.
- *Base de conocimiento.* Conjunto de reglas que se activan a partir de los hechos para producir ciertos resultados.
- *Motor de inferencia.* Permite controlar la ejecución de las reglas.

ClipsPy involucra las capacidades de CLIPS dentro del ecosistema de Python.

## A.1 Instalación

### En Windows

Ingrese el siguiente comando:

```
pip install clipspy==0.3.2
```

### En Linux

Ingrese el siguiente comando:

```
sudo pip install clipspy==0.3.3
```

## A.2 Requerimientos

**Python** - 2.7,- 3.4,- 3.5,- 3.6,- 3.7

**C++** 9.0

## A.3 Enlaces

**Código Fuente** $https://github.com/noxdafox/clipspy$

**Documentación** $https://clipspy.readthedocs.io$

**Descarga** $https://pypi.org/project/clipspy/$

## A.4 Ejemplo

En el Listado A.1 se hace uso de clips desde Python. Se inicia definiendo el entorno de Clips, posteriormente se carga un archivo llamado "*rules.clp*" que contiene las reglas que nos permitirán evaluar los hechos ingresados.

En el código se hace uso de un archivo "*rules.clp*", este se encuentra en el listado A.2

Listado A.1: Ejemplo básico del uso de clipspy en Python

```
"""
  Ejemplo basico de uso de Clipspy
"""
import clips


env = clips.Environment()
rules_file = open('rules.clp', 'r')
rules = rules_file.read()
rules_file.close()


env.build(rules)


fact = ''
while True:
  fact = input('new␣fact␣or␣(c)␣to␣continue:␣')
  if fact == 'c':
    break
  env.load_facts(fact)
```

```
  env.run()


for fact in env.facts():
  print(fact)
```

Este listado (Listado A.2) cuenta con dos reglas que permiten evaluar los hechos ingresados. La primera regla (rule_00) indica que se deben cumplir dos hechos: 1) que la entidad a evaluar se encuentre en la base de datos, 2) que la pila en la que se colocarán los elementos se encuentre vacía, si estos dos se cumplen la regla es activada.

La segunda regla (rule_01) se activa cuando se cumplen tres hechos: 1) la entidad a evaluar se encuentra en la base de datos, 2) la entidad que se va a evaluar es más pequeña (en cuanto al tipo) que la entidad que se encuentra en el top de la pila, 3) existe una asociación entre la entidad a evaluar y la entidad que se encuentra en el top de la pila. De ser así la regla es activada.

Listado A.2: Archivo de reglas

```
(defrule rule_00 "Empty -->Guerrero"
  ?fact_1 <- (in_db ?entity yes)
  ?fact_2 <-(stack is_empty yes)


=>
  (printout t ?entity ".setHighestFeatureCode()" crlf)
  (printout t "stack.push("?entity ")" crlf)
  (retract ?fact_1)
  (retract ?fact_2)
  (assert (stack is_empty no))
  (assert (undetermined no))
)
```

```
(defrule rule_01 "Guerrero -->Tlacoachistlahuaca"
  ?fact_1 <- (in_db ?entity yes)
  ?fact_2 <- (predecessor ?entity ?top yes)
  ?fact_3 <- (association_between ?entity ?top yes)
  (stack is_empty no)
  (undetermined no)
=>
    (retract ?fact_1)
  (retract ?fact_2)
  (retract ?fact_3)
  (printout t "stack.push("?entity ")" crlf)

)
```

En el listado A.3 se observa un ejemplo de hechos posibles a ingresar en los que se cumplirían las dos reglas definidas en el listado A.2.

Listado A.3: Ejemplo de hechos

```
(in_db Guerrero yes)
(stack is_empty yes)
(in_db Tlacoachistlahuaca yes)
(predecessor Tlacoachistlahuaca Guerrero yes)
(association_between Tlacoachistlahuaca Guerrero yes)
```

# B

# Ejemplo de aplicación

Para ilustrar mejor el funcionamiento del método propuesto, se presenta la Figura B.1, en la cual se muestra el texto inicial, las reglas activadas con cada topónimo ingresado y la pila de desambiguación con el estado después de procesar todos los topónimos.

En la parte superior de la imagen se encuentra el texto de entrada. Como se puede observar, cada topónimo se resalta en negritas y con ciertos colores. Estos colores se explican en el diagrama del lado izquierdo, en este se ejemplifica la jerarquía entre topónimos. Como se puede observar, de color morado se encuentra la entidad de más alto nivel jerárquico, en este caso, el país México, este se incluye en el diagrama a pesar de no expresarse directamente en el texto. Los demás grupos de colores representan un código distinto explicado en la Tabla 3.2.

Debajo del texto, del lado derecho se encuentra la lista de reglas activadas para cada topónimo presente en el texto. Las reglas definidas se describen en la Tabla 3.3. La activación de cualquier regla implica cambios sobre la pila de desambiguación. Esta pila se encuentra debajo de la lista de reglas, en ésta se muestra el estado final de la pila.

Las acciones siguientes representan un seguimiento paso a paso del comportamiento del método propuesto con respecto al ejemplo de la Figura B.1.

1. El topónimo *Mueblería Tu Hogar* no se encuentra en el gazetteer ($\mathbb{G}$), además debido a que la pila de desambiguación ($\mathbb{S}$) se encuentra vacía, se activa la regla $R_5$. La cual indica que el topónimo será colocado en la pila de conflictos ($\mathbb{C}$). La pila de conflictos es la encargada de almacenar los topónimos con coordenadas desconocidas hasta que es posible asociarlos con otros.

2. El topónimo *Ciudad de México* se encuentra en $\mathbb{G}$, además la pila $\mathbb{S}$ todavía está vacía. Esto provoca la activación de la regla $R_0$, lo cual lleva a colocar este topónimo en $\mathbb{S}$.

3. El topónimo *Azcapotzalco* se encuentra en $\mathbb{G}$, además $\mathbb{S}$ no se encuentra vacía, por lo que la regla $R_1$ es activada, lo cual indica que Azcapotzalco será colocado en el top de $\mathbb{S}$ asociada a la entidad anterior (*Ciudad de México*).

4. El topónimo *Ixcatán* se encuentra en $\mathbb{G}$, $\mathbb{S}$ no está vacía y este topónimo no está asociado al anterior. Estos hechos activan la regla $R_2$, la cual nos indica que *Ixcatán* será colocada sin asociación con las anteriores.

5. Los hechos activados por el siguiente topónimo, *Zapopan*, provocan la activación de la regla $R_3$. Esta regla realiza un cambio en la pila y permite colocar este nuevo topónimo asociado al antepenúltimo topónimo en $\mathbb{S}$ (*Azcapotzalco*) sin asociarla a *Ixcatán*. Hasta el momento en el top de la pila se encuentra *Ixcatán*, debajo de ésta *Azcapotzalco* y en la base la pila *Ciudad de México*.

6. Al igual que en el paso anterior, *Jalisco* activa la regla $R_3$, colocando nuevamente *Ixcatan* en el top y *Jalisco* debajo de ésta, pero en esta ocasión *Ixcatan* y *Jalisco* si están asociadas.

7. El topónimo *Pedregal de Santo Domingo* activa la regla $R_6$, lo cual indica que este topónimo será colocado en el top de $\mathbb{S}$ sin relación con el topónimo anterior.

8. El topónimo *Coyoacán* activa la regla $R_3$, lo que realiza un cambio en la pila, colocando este topónimo debajo del top actual.

9. El topónimo *Zapatería Juárez* no se encuentra en $\mathbb{G}$, lo que activa la regla $R_4$.

10. Como paso final se hace una revisión de $\mathbb{C}$ para conocer su estado, como en este caso *Mueblería tu hogar* se encuentra ahí, se activa la regla $R_7$.

En la Figura B.2 se muestran los topónimos con sus coordenadas geográficas en un mapa. Este mapa es el que devuelve el sistema a partir del estado final de la pila de desambiguación $\mathbb{S}$. Se encuentra divido en dos partes, en la parte superior se ve la parte correspondiente al *estado de Jalisco*, con sus respectivos topónimos relacionados a este estado. En la parte inferior se encuentra lo relacionado a la *Ciudad de México*. Cada uno de estos topónimos cuenta con dos pines por entidad, el color azul es la opción que el método considera correcta, el color negro indica una ubicación alterna para este topónimo. De acuerdo al ejemplo de la Figura B.1, se identificaron dos topónimos que no estaban en $\mathbb{G}$, por lo que no contaban con coordenadas propias, la forma en la que el método propuesto soluciona esto es asignándolo a la entidad más cercana a ella en la pila $\mathbb{S}$, es por esto que en la Figura B.2 *Zapatería Juárez* comparte las coordenadas de *Pedregal de Santo Domingo* y *Mueblería Tu Hogar* comparte las coordenadas de Azcapotzalco.

Llega un cliente a la **Mueblería Tu Hogar**, en la **Ciudad de México** sucursal **Azcapotzalco**. Solicita un comedor que está en oferta, pero pide que se realice un envio a domicilio al poblado de **Ixcatán** en el municipio de **Zapopan**, **Jalisco**.
El problema es que la sucursal encargada de hacer los envíos a esa zona es la de **Pedregal de Santo Domingo**, en el municipio de **Coyoacán**, junto a la **Zapatería Juárez**, pero en esa sucursal el comedor no tiene descuento. La mueblería deberá llegar a un acuerdo con el cliente.

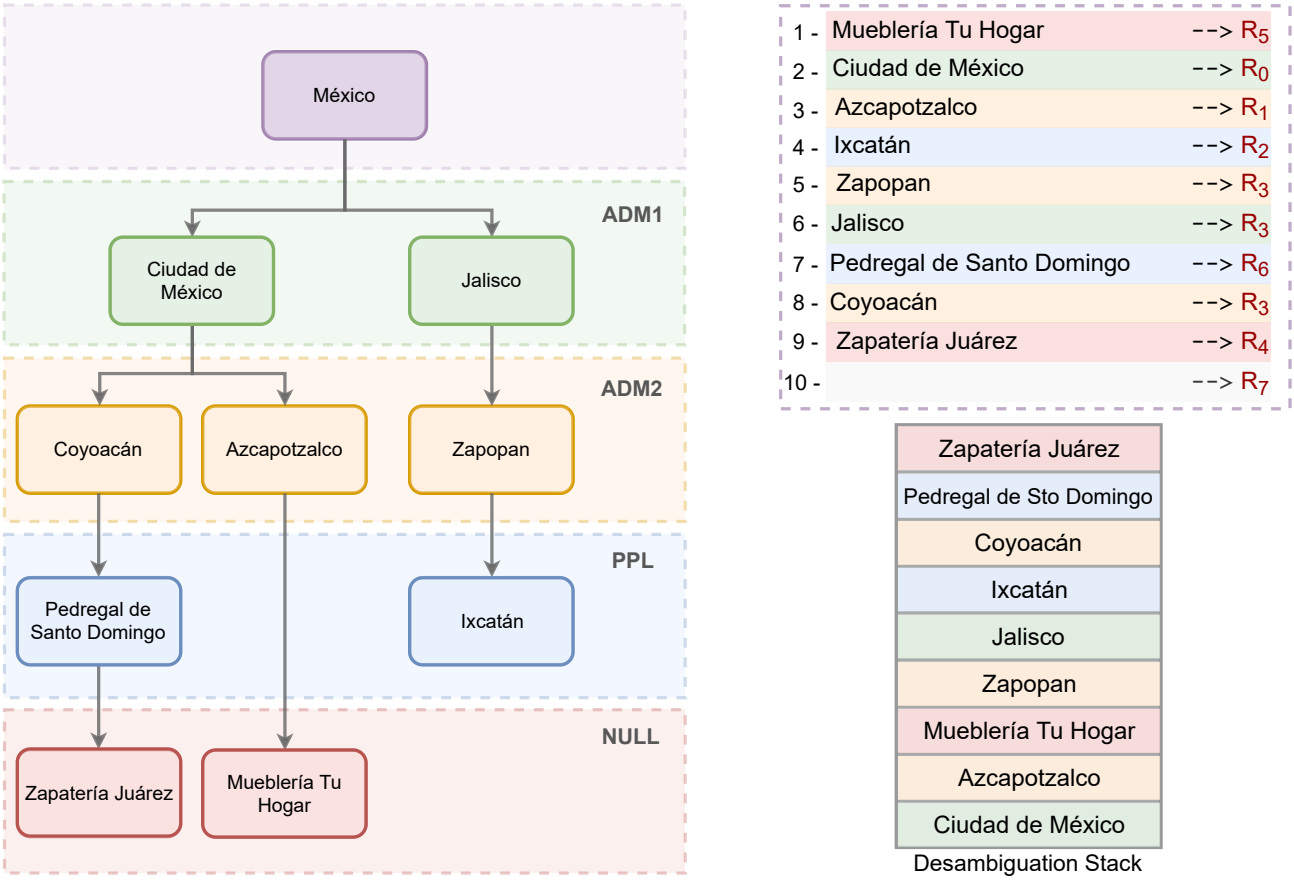


Figura B.1: Ejemplo de uso del método propuesto

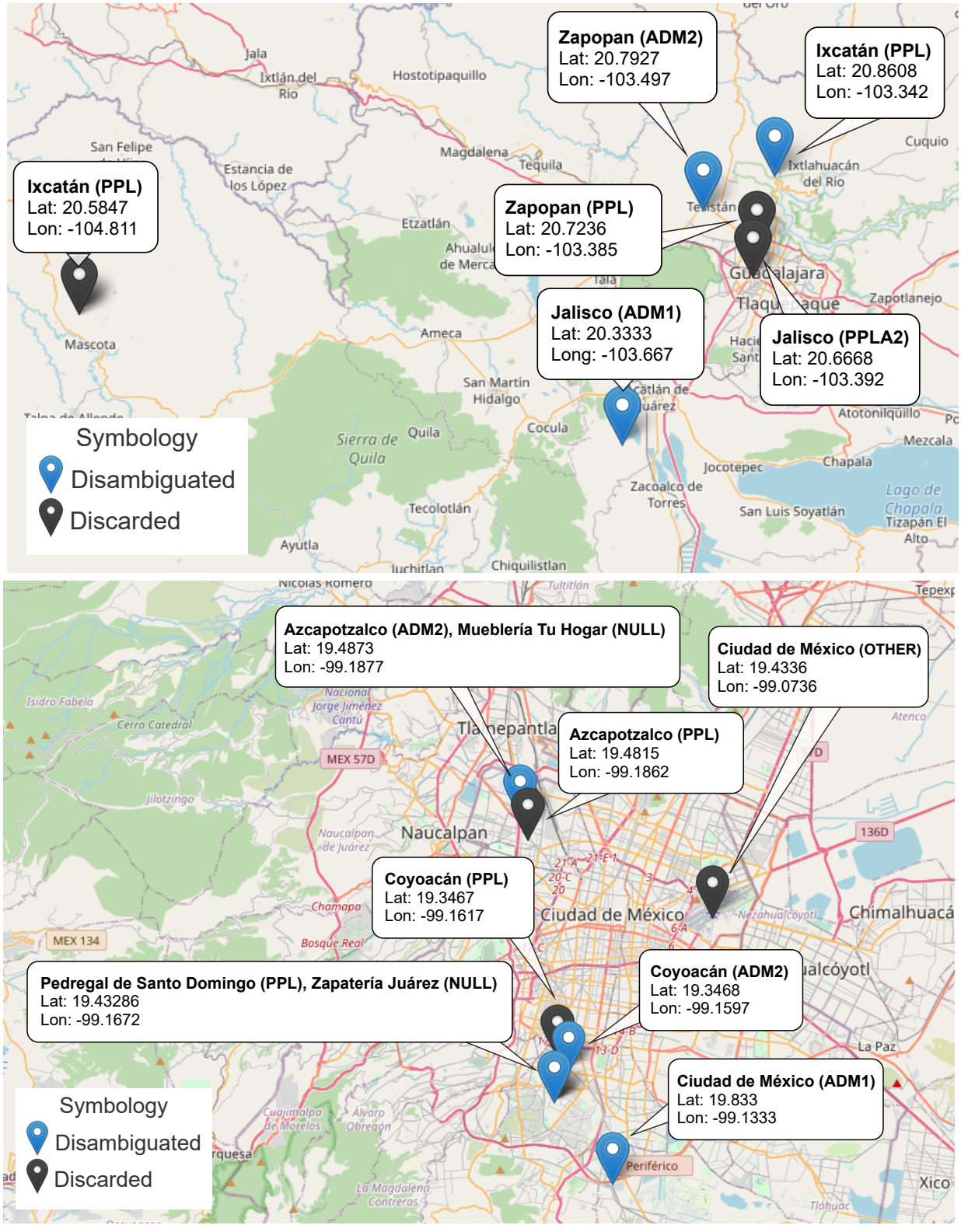


Figura B.2: Mapa obtenido a partir del ejemplo de la Figura B.1

# Bibliografía


[1] Abelleira, M. A. P. and Cardoso, C. A. (2010). Minería de texto para la categorización automática de documentos. *Facultad de Ingeniería e Informática e IESIING.*

[2] Alexopoulos, P., Ruiz, C., and Gomez-Perez, J. (2012). Optimizing geographical entity and scope resolution in texts using non-geographical semantic information. In *Proceedings of the 6th International Conference on Advances in Semantic Processing, SEMAPRO*, pages 65 – 70.

[3] Alexopoulos, P., Ruiz, C., Villazon-Terrazas, B., and Gomez-Perez, J.-M. (2013). Klocator: an ontology-based framework for scenario-driven geographical scope resolution. *Int. J. Adv. Intell. Syst*, 6(3 - 4):177 – 187.

[4] Amitay, E., Har'El, N., Sivan, R., and Soffer, A. (2004). Web-a-where: Geotagging web content. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 273 – 280, New York, NY, USA. ACM.

[5] Andogah, G., Bouma, G., and Nerbonne, J. (2012). Every document has a geographical scope. *Data and Knowledge Engineering*, 81 - 82:1 – 20.

[6] Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley, Harlow.

[7] Barrera, M. C. (2015). Minería de texto: una visión actual. *Biblioteca Universitaria*, 17(2):129–138.

[8] Behrens, C. A. and Bassu, D. (2006). Information retrieval and text mining using distributed latent semantic indexing. US Patent 7,152,065.

[9] Brusa and Konstantinides (2012). Procedimiento para la georreferenciación de las parcelas del catastro territorial de la provincial de córdoba. Technical Report 2, Escuela de Agrimensura. Universidad Nacional de Córdoba.

[10] Cardoso, N. (2011). Evaluating geographic information retrieval. *SIGSPATIAL Special*, 3(2):46 – 53.

[11] Clough, P., Sanderson, M., and Joho, H. (2004). Extraction of semantic annotations from textual web pages. *Deliverable D15*, 6201.

[12] Cowie, J. and Lehnert, W. (1996). Information extraction. *Commun. Association for Computing Machinery*, 39(1):80–91.

[13] D. Manning, C., Raghavan, P., and Schütze, H. (2009). *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, England.

[14] Dang, S. and Ahmad, P. A. (2014). Text mining: Techniques and its application. *International Journal of Engineering & Technology Innovations,*, 1.

[15] de la Torre, J. and del Consuelo, M. (2017). *Nuevas técnicas de minería de textos: Aplicaciones*. PhD thesis, Universidad de Granada.

[16] Delboni, T., Borges, K. A., Laender, A. H., and Davis Jr, C. A. (2007). Semantic expansion of geographic web queries based on natural language positioning expressions. *Transactions in GIS*, 11(3):377 – 397.

[17] Goodchild, M. F. and Hill, L. L. (2008). Introduction to digital gazetteer research. *International Journal of Geographical Information Science*, 22(10):1039 – 1044.

[18] Goyal, A., Gupta, V., and Kumar, M. (2018). Recent named entity recognition and classification techniques: a systematic review. *Computer Science Review*, 29:21 – 43.

[19] Greengrass, E. (2000). Information retrieval: A survey.

[20] Gritta, M., Pilehvar, M. T., Limsopatham, N., and Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, 52(2):603–623.

[21] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199 – 220.

[22] Gudivada, V. (2018). *Natural Language Core Tasks and Applications*, pages 403–428. North-Holland.

[23] Gupta, R. (2014). *Conditional Random Fields*, pages 146 – 146. Springer US, Boston, MA.

[24] Hariharan, R., Hore, B., Li, C., and Mehrotra, S. (2007). Processing spatial-keyword (sk) queries in geographic information retrieval (gir) systems. In *19th International Conference on Scientific and Statistical Database Management (SSDBM 2007)*, page 16.

[25] Hill, L. L. (2000). Core elements of digital gazetteers: Placenames, categories, and footprints. In Borbinha, J. and Baker, T., editors, *Research and Advanced Technology for Digital Libraries*, pages 280 – 290, Berlin, Heidelberg. Springer Berlin Heidelberg.

[26] Inkpen, D., Liu, J., Farzindar, A., Kazemi, F., and Ghazi, D. (2017). Location detection and disambiguation from twitter messages. *Journal of Intelligent Information Systems*, 49(2):237 – 253.

[27] Khalid, M. A., Jijkoun, V., and de Rijke, M. (2008). The impact of named entity normalization on information retrieval for question answering. In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., and White, R. W., editors, *Advances in Information Retrieval*, pages 705 – 710, Berlin, Heidelberg. Springer Berlin Heidelberg.

[28] Konkol, M. and Konopik, M. (2015). Segment representations in named entity recognition. In Kral, P. and Matousek, V., editors, *Text, Speech, and Dialogue*, pages 61 – 70, Cham. Springer International Publishing.

[29] Lieberman, M. D., Samet, H., and Sankaranarayanan, J. (2010). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th International Conference on Data Engineering (ICDE 2010)*, pages 201 – 212.

[30] Mani, I., Hitzeman, J., Richer, J., Harris, D., Quimby, R., and Wellner, B. (2008). Spatialml: Annotation scheme, corpora, and tools.

[31] Mariano, C. and Hernan, M. (2012). Verificación y densificación de la red de catastro de la provincia de córdoba. Technical Report 2, Escuela de Agrimensura. Universidad Nacional de Córdoba.

[32] Martins, B. and Silva, M. (2005). A graph-ranking algorithm for geo-referencing documents. In *Fifth IEEE International Conference on Data Mining*, pages 741–744.

[33] Mikheev, A., Moens, M., and Grover, C. (1999). Named entity recognition without gazetteers. In *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics*, EACL '99, pages 1 – 8, Stroudsburg, PA, USA. Association for Computational Linguistics.

[34] Molina-Villegas, A., Siordia, O. S., Aldana-Bobadilla, E., Aguilar, C. A., and Acosta, O. (2019). Extracción automática de referencias geoespaciales en discurso libre usando técnicas de procesamiento de lenguaje natural y teoría de la accesibilidad. *Procesamiento del Lenguaje Natural*, 63:143 – 146.

[35] Monteiro, B., Davis, C., and Fonseca, F. (2016). A survey on the geographic scope of textual documents. *Computers & Geosciences*, 96:23 – 34.

[36] Montes Gómez, M., Gelbukh, A., and López López, A. (2005). Minería de texto empleando la semejanza entre estructuras semánticas. *Computación y Sistemas*, 9:63 – 81.

[37] Mostern, R. and Johnson, I. (2008). From named place to naming event: creating gazetteers for history. *International Journal of Geographical Information Science*, 22(10):1091– 1108.

[38] Navigli, R. (2009). Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):10.

[39] Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab.

[40] Pantaleo, G. and Nesi, P. (2014). Ge(o)lo(cator): Geographic information extraction from unstructured text data and web documents. In *2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization*, pages 60 – 65.

[41] Powers, D. M. (1984). Natural language the natural way. *Computer Compacts*, 2(3):100 – 109.

[42] Pérez-Rodríguez, R., Anido-Rifón, L., Gómez-Carballa, M., and Mouriño-García, M. (2016). Architecture of a concept-based information retrieval system for educational resources. *Science of Computer Programming*, 129:72 – 91. Special issue on eLearning Software Architectures.

[43] Radke, M., Gautam, N., Tambi, A., Deshpande, U., and Syed, Z. (2018). Geotagging text data on the web a geometrical approach. *IEEE Access*, 06:30086–30099.

[44] Riley, G. (1991). Clips: An expert system building tool. *In NASA, Washington, Technology 2001: The Second National Technology Transfer Conference and Exposition*, 2:149 – 158.

[45] Rupp, C., Rayson, P., Baron, A., Donaldson, C., Gregory, I., Hardie, A., and Murrieta-Flores, P. (2013). Customising geoparsing and georeferencing for historical texts. In *Proceedings of the IEEE International Conference on Big Data, Big Data*, pages 59 – 62.

[46] Sanderson, M. (1994). Word sense disambiguation and information retrieval. In *SIGIR'94*, pages 142–151. Springer.

[47] Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition.

[48] Senapati, A., Das, A., and Garain, U. (2007). Named-entity recognition in bengali. In *Post-Proceedings of the 4th and 5th Workshops of the Forum for Information Retrieval Evaluation*, pages 14:1 – 14:5, New York, NY, USA. ACM.

[49] Silva, M. J., Martins, B., Chaves, M., Afonso, A. P., and Nuno, C. (2006). Adding geographic scopes to web resources. *Computers, Environment and Urban Systems*, 30(4):378 – 399.

[50] Tobin, R., Grover, C., Byrne, K., Reid, J., and Walsh, J. (2010). Evaluation of georeferencing. In *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR 10, pages 1 – 8, New York, NY, USA. ACM.

[51] Torruella, J. and Llisterri, J. (1999). Diseno de corpus textuales y orales. *Filologia e informatica. Nuevas tecnologias en los estudios filologicos*, pages 45 – 77.

[52] Vaid, S., Jones, C. B., Joho, H., and Sanderson, M. (2005). Spatio-textual indexing for geographical search on the web. In Bauzer Medeiros, C., Egenhofer, M. J., and Bertino, E., editors, *Advances in Spatial and Temporal Databases*, pages 218 – 235, Berlin, Heidelberg. Springer Berlin Heidelberg.

[53] Vasardani, M., Winter, S., and Richter, K.-F. (2013). Locating place names from place descriptions. *International Journal of Geographical Information Science*, 27(12):2509 – 2532.

[54] Zheng, W. and Blake, C. (2010). Bootstrapping location relations from text. *Proceedings of the American Society for Information Science and Technology*, 47:1 – 9.

[55] Zitouni, I. (2014). *Natural language processing of semitic languages*. Springer.

[56] Zubizarreta, A., de la Fuente, P., Cantera, J. M., Arias, M., Cabrero, J., García, G., Llamas, C., and Vegas, J. (2008). A georeferencing multistage method for locating geographic context in web search. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 1485–1486, New York, NY, USA. ACM.