Geontomex: una ontología espacial de México para la desambiguación de topónimos

Belém Priego Sánchez, María J. Somodevilla García, Ivo H. Pineda Torres y Jaime Hernández Gómez

Benemérita Universidad Autónoma de Puebla,
Av. San Claudio 14 Sur, Col. San Manuel Puebla, México
{belemps, mariajsomodevilla, ivopinedatorres, kngjaime }@gmail.com

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

Palabras clave: Ontología, recuperación de información geográfica, relaciones espaciales, opónimo, desambiguación.

1. Introducción

La desambiguación de topónimos (DT) es una de las tareas de la recuperación de información (Information Retrieval, IR), más específicamente en la recuperación de información geográfica (Geographical Information Retrieval, GIR) y búsqueda de respuestas (Question Answering, QA), incluyendo además, la generación de mapas. DT tiene como objetivo relacionar nombres de lugares con su representación geográfica. La Geo-información es abundante en la Web y bibliotecas digitales, por ejemplo: colecciones de fotografías geo-referenciadas (Flickr), noticias, y bases de datos de información demográfica (En México a cargo del Instituto Nacional de Estadística, Geografía e Informática, INEGI). Aproximadamente el 80% de las páginas web contienen referencias a lugares [1], mucha de la información necesaria está relacionada a un contexto geográfico dado, por ejemplo: encontrar los restaurantes más *cercanos*, encontrar noticias acerca de *México*, así como encontrar fotografías tomadas en *Cholula*, entre otras. Se ha reportado también en la bibliografía, que aproximadamente el 20% de las consultas en la Web son de naturaleza local [2]; esto quiere decir que la información geográfica está omnipresente.

GIR pertenece a una rama de la recuperación de información, que incluye todas las tareas de investigación que tradicionalmente forman el núcleo de la IR, pero además con un énfasis en la información geográfica y espacial. En la tarea de GIR, la mayoría de las peticiones de los usuarios son del tipo *"X" en "P"* donde *P* representa un nombre de lugar y *X*, la parte temática de la consulta [14]. GIR aborda dificultades de IR, tales como: ambigüedad Geográfica (topónimos) por ejemplo: existe una *catedral* en *St. Paul* en *Londres* y otra en *Sao Paulo*, o el caso de regiones geográficas mal definidas: *"cerca del este"*, regiones geográficas complejas: *"cerca de ciudades rusas"* o *"a lo largo de la costa mediterránea"*, aspectos multilingües como son: *"GreaterLisbon"* en inglés significa lo mismo que *"Grande Lisboa"* en portugués o que *"GroBraumLissabon"* en alemán y la existencia de la granularidad en las referencias a países: *"al norte de Italia"*.

La información puede ser geo-referenciada por nombres de lugares o por coordenadas geográficas. Para este caso tenemos las siguientes clasificaciones: formal e informal y mediante un recurso (repositorio de sentidos). El repositorio de sentidos hace referencia a la utilización de diccionarios, tesauros u ontologías, en el cual se indiquen los distintos sentidos de las palabras [14].

Existen diferentes ontologías de topónimos reportadas en la bibliografía tales como Geo-WordNet [11], GeoNames [12], TRIPOD [13], por citar algunas, pero la mayoría de estas incluyen topónimos del lenguaje inglés. También se han comenzado a reportar algunas ontologías de topónimos para los lenguajes alemán [15], portugués [16], chino [17], entre otras. Sin embargo, no se han encontrado reportados en la bibliografía trabajos relacionados con ontologías de topónimos en español. Por lo tanto se hace imperativo el desarrollo de ontologías tales como la que proponemos, para soportar la tarea de resolución de topónimos presentes en los documentos escritos en español que se recuperan por peticiones en la Web.

En este trabajo se propone una ontología espacial de México para la tarea de desambiguación de topónimos. Una ontología espacial es una estructura que permite describir adecuadamente las características específicas del espacio geográfico. Una ontología proporciona un vocabulario de clases y relaciones para describir un ámbito determinado; éstas permiten que las máquinas puedan intercambiar información de forma efectiva y eficiente. Para ello proporcionan formalismos y estructuran la información permitiendo un cierto grado de razonamiento automático.

El resto del artículo está estructurado en las siguientes secciones; en la Sección 2 se explica la tarea de desambiguación de topónimos. El repositorio de sentidos se discute en la Sección 3, con el propósito de sentar las bases para la descripción de la propuesta presentada en la Sección 4. Finalmente, en la Sección 5 se reportan las conclusiones y se presenta el posible trabajo a futuro.

2. Desambiguación de topónimos

Los topónimos pueden ser ambiguos y además tener algún tipo de ambigüedad ya sea GEO/GEO o GEO/NO-GEO. En el caso de la ambigüedad de topónimos GEO/GEO un topónimo representa varios lugares geográficos, por ejemplo, *"Trípoli"* que es el nombre de 16 lugares en el mundo. En el caso de la ambigüedad

de topónimos GEO/NO-GEO, un topónimo puede referirse a entidades geográficas y no geográficas, por ejemplo, *"Benito Juárez"* representa a una persona y también a lugares; otro ejemplo, *"Java"* es una isla indonesia y a su vez un lenguaje de programación. Para ambos tipos de ambigüedad, los dominios de aplicación son la extracción y recuperación de información.

La desambiguación de topónimos constituye una de las tareas importantes dentro de la recuperación de información geográfica. Recientemente, ha habido un gran interés en el problema de desambiguación de topónimos desde distintas perspectivas como el desarrollo de recursos para la evaluación de los métodos de desambiguación de topónimos [3] y el uso de estos métodos para mejorar la resolución del alcance (*scope*) geográfico en documentos electrónicos [4], entre los trabajos más representativos. No sería posible estudiar la ambigüedad de los topónimos sin estudiar también los recursos que se involucran en el proceso, como bases de datos, diccionarios y otros que se usan para encontrar los significados diferentes de una palabra. Considerando que los métodos para la desambiguación de topónimos son de diferente naturaleza, estos aún tienen factores en común. La mayoría están influenciados por dos fases principales, la primera fase es extraer los referentes candidatos y la segunda fase es seleccionar el referente correcto.

Dada la importancia del tema desarrollado y una vez realizada la revisión bibliografía se encontraron muy pocos trabajos sobre desambiguación de topónimos para el idioma español y eso precisamente fue lo que motivó el desarrollo del presente trabajo. En la siguiente sección se describe el repositorio de sentidos: la ontología espacial, que describe formalmente los distintos sentidos de los objetos geográficos naturales y artificiales en base a los conceptos y relaciones del dominio.

3. Ontologías

Las ontologías permiten que las máquinas puedan intercambiar información de forma efectiva y eficiente. Para ello proporcionan formalismos y estructuran la información permitiendo un cierto grado de razonamiento automático. Gruber [5] creó una de las definiciones más citadas del concepto de ontología en el ámbito de la informática: "una especificación explícita y formal sobre una conceptualización compartida".

Gruber [6] es también uno de los autores más citados al identificar los cinco componentes básicos del modelado de ontologías, los cuales se enuncian a continuación:

- Conceptos. Son las ideas básicas que se intentan formalizar.
- Relaciones. Representan las interacciones y los enlaces entre los conceptos del dominio.
- Funciones. Son casos especiales de relaciones donde se identifican elementos mediante el cálculo de una función que considera varios elementos de la ontología.
- Instancias. Se usan para representar elementos determinados en una ontología.

- Axiomas. Los axiomas formales sirven para modelar sentencias que son siempre ciertas. Normalmente, se usan para representar conocimiento que no puede ser formalmente definido por los componentes descritos anteriormente consiguiendo así una mayor capacidad expresiva del dominio. Además, también se usan para verificar la consistencia de la propia ontología.

3.1. Ontologías espaciales

Las Ontologías Espaciales, de acuerdo a [7], son una extensión de la Lógica Descriptiva (DL), con un dominio concreto para la dimensión espacial (es decir, considera objetos espaciales tales como puntos, líneas, polígonos), para así permitir la combinación de representación del conocimiento y el razonamiento espacial dentro de un paradigma único. El dominio concreto está definido por un conjunto de predicados representando relaciones topológicas entre objetos. La habilidad de definir roles topológicos facilita la especificación de conceptos y objetos espaciales. También provee acceso algoritmos de razonamiento espacial que permiten la extensión del razonamiento terminológico de la dimensión espacial.

Algunos investigadores, como Spaccapietra [8], dividen esta Ontología en espacio y tiempo; las ontologías de tiempo definen los conceptos que son usados en un tiempo especificado y elementos temporales, como son las instancias, intervalos, cronómetros, entre otros, y relaciones temporales como son precedentes, antecedentes, entre los más significativos, pero de igual manera son denominados espaciales. Debido a la aparición de una gran cantidad de información geográfica y mapas en Internet, de casi todos los sitios posibles sobre la tierra, aparecen los Servicios de la Web Semántica [9] que convienen a un tipo de tecnologías más elaboradas, en un mundo donde se cree que más del 80% de los datos tiene un componente geográfico, como lo son las nuevas aplicaciones de mapas publicadas en el Web. Los mapas web muestran recientemente grandes crecimientos, su integración dentro del dominio espacial aparece como un paso esencial hacia la adopción de la tecnología SWS (Shore Wireless Service). Sin embargo, el espacio geográfico como un único pero total dominio encuadrado tiene especificaciones que describen semánticas más reconocidas. Además, los Sistemas de Información Geográfica (GIS) necesitan adoptar habilidades humanas cognitivas de representación espacial y razonamiento.

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

4. Geontomex: desarrollo y descripción

El desarrollo de la ontología geográfica propuesta en este artículo, se enmarca en un proyecto global ambicioso, cuyo objetivo es de ser tomado como repositorio de sentidos para resolver la tarea de desambiguación topónimos. Cabe mencionar que esta ontología puede ser utilizada para desambiguar topónimos en consultas a la Web, relacionadas por ejemplo con: el ejercicio y la promoción del turismo en México,

aplicaciones de cambio climático, realización de planeaciones urbanas y desarrollo de planes estratégicos de tipo económico-sociales entre otras. Dentro de este proyecto, uno de los aspectos a desarrollar es la implementación de una ontología que permita identificar el topónimo correcto en un contexto (*corpus)* y a su vez mediante la ontología obtener su posición geográfica (latitud y longitud).

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

Existen varios lenguajes ontológicos para implementar ontologías, los cuales proporcionan distintos niveles de formalismo y facilidad de razonamiento. El lenguaje OWL[1] (Ontology Web Language), estandarizado por el W3C (World Wide Web Consortium), permite definir ontologías con varios niveles de detalle. Dicho lenguaje se puede categorizar en tres especies o sublenguajes: OWL-Lite, OWL-DL y OWL-Full. La ontología espacial se implementó en Protégé[2] empleando el sublenguaje OWL-DL, debido a que está diseñado para aquellos usuarios que requieren máxima expresividad conservando completitud computacional (se garantiza que todas las conclusiones sean computables) y resolubilidad (todos los cálculos se resolverán en un tiempo finito) [10]. Una de las ventajas de utilizar este Protégé es que cuenta con el manejo de instancias sobre las clases, así como restricciones para generar éstas.

4.1. Desarrollo: Geontomex

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

Para la definición de aspectos genéricos de la ontología, han servido como base la ontología presentada en [10], una ontología mixta que proporciona un vocabulario de clases y relaciones para describir un área específica. En este caso el espacio geográfico incluye un análisis de la distribución de 500 millones de hispanohablantes estimados en el mundo.

Conceptos (Jerarquía de Clases)

La ontología se desarrolla a partir de la jerarquía de clases que se muestran en la figura 1. En esta figura se pueden distinguir tres clases de alto nivel: *Extension_Geografica, Estructura_Geografica* y *Localizacion*.

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

[1] www.w3.org/TR/owl-features

[2] [2]protege.stanford.edu

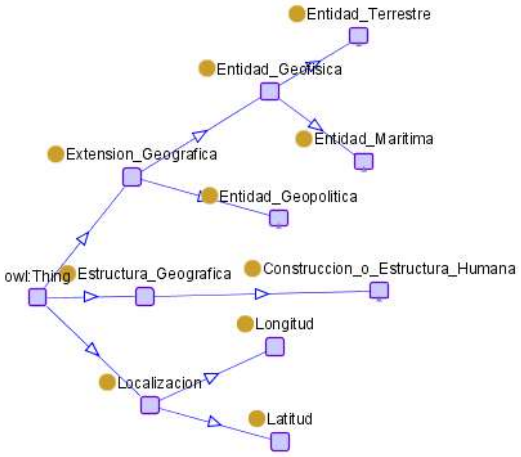

Figura 1. Taxonomía de la Ontología Espacial: Geontomex

- *Extructura_Geografica*: representa la clasificación más habitual de áreas o espacios artificiales creados por el ser humano. Cuenta con una subclase: Construccion_o_Estructura_Humana que a su vez tiene diez subclases: *Aeropuerto, Calle_o_Carretera, Camino, Canal, Ferrocarril, Monumento, Parque_Natural, Presa, Puente* y *Puerto.*
- *Extension_Geografica*: representa los espacios geográficos naturales; dentro de esta clase se incluyen las subclases: *Entidad_Geofisica*, que define los espacios naturales marítimos y terrestres de México y l*a Entidad_Geopolitica*, que representa la división política de México.
- *Localizacion*: representa una manera de definir la localización de un lugar.

Relaciones

- **Propiedades de objeto**

A partir de la jerarquía de clases presentada en la figura 1, se definen una serie de propiedades de objeto para, principalmente relacionar todas las clases de la ontología con la entidad *Localizacion*. A partir de las relaciones, es posible definir todos los aspectos que se quieren relacionar con la localización de los topónimos. A continuación en la Tabla 1 se detallan las relaciones descritas como propiedades de objeto incluidas en la ontología.

- **Propiedades de datos**

Además de las propiedades que relacionan las diferentes entidades de la ontología, es necesario crear propiedades de datos que describen dichas entidades. Las principales en esta ontología son las que describen la entidad *Localizacion* y son las siguientes *latitud* y *longitud*. Estas son consideradas propiedades funcionales ya que describen los valores de la localización de un topónimo.

Tabla 1. Propiedades de objeto de Geontomex.

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

Instancias

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

Para la tarea de desambiguación de topónimos se está utilizando como contexto un Corpus de noticias multilingüe (incluye los idiomas: español, inglés, francés, italiano y portugués) de la tarea de búsqueda de respuestas (QA) de la iniciativa CLEF[3], con una colección total de 4264 documentos de los años 2003 a 2005. Para la tarea que se

[3] http://www.clef-initiative.eu//

está realizando sólo se ocupó el corpus en el idioma español, el cual consta de 731 documentos. El corpus anterior comprende 216102 noticias de todo el mundo y cada una contiene topónimos.

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

Axiomas

Los axiomas están en desarrollo, teniendo como principal axioma de Geontomex la relación espacial "*cerca*", esta relación se está diseñando y se tiene previsto incorporarla para seguir verificando la consistencia de la ontología. Este axioma trata el concepto de ambigüedad espacial y se está programando con las relaciones básicas de la ontología que son*: is_part_of* e *is_a,* teniendo como base la siguiente expresión:

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

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

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

4.2. Validación de Geontomex

Para garantizar la consistencia de Geontomex, ésta se validó utilizando un razonador espacial. En este proceso de validación, la información acerca de los objetos en el espacio y sus interrelaciones son recogidas por varios medios, tales como medidas, observaciones, o inferencia, y se utilizan para llegar a conclusiones válidas conforme a las relaciones de objeto o para determinar la forma de realizar una tarea. El razonamiento espacial es usado para inferir todas las relaciones posibles entre un conjunto de objetos usando un subconjunto de las relaciones especificadas. RacerPro[4] es un razonador utilizado tanto para Lógica Descriptiva Básica, como para muy expresiva y espacial, por este motivo Geontomex fue validada mediante este razonador. Además, también puede ser usado como un sistema para gestionar las ontologías de la Web Semántica basadas sobre OWL, es decir, puede ser usado como un motor para editores ontológicos como Protégé.

[4] Sitio oficial RacerPro: http://www.racer-systems.com/

El editor de RacerPro utilizado para la validación, fue RacerPorter[5] en el cual se pueden cargar bases de conocimiento, conmutar entre diferentes taxonomías, inspeccionar las instancias, visualizar TBoxes y ABoxes, manipular los servicios, entre otros servicios que la interfaz maneja, para la validación, razonamiento espacial y verificación de consistencia de ontologías.

5. Conclusión y trabajo futuro

En este artículo se presentó Geontomex, una ontología espacial que sirve como repositorio de sentidos para una de las tareas importantes de la recuperación de información geográfica como lo es la desambiguación de topónimos para el idioma español en particular para la República Mexicana; cabe resaltar que existen pocas investigaciones sobre esta temática para el idioma español en comparación con otros idiomas, en específico con el idioma inglés.

Se está trabajando en el enriquecimiento del corpus para ser utilizado como contexto a través de la utilización de técnicas de bootstrapping. El corpus enriquecido se utilizará entonces como conjunto de prueba para generar un modelo de clasificación de topónimos tipo GEO/GEO Y GEO/NOGEO, el cual se plantea como meta final de este proyecto de investigación.

Referencias

1. Spink, A., Wolfram, D., Jansen, M.B.J., Saracevic, T.: Searching the web: the public and their queries. Journal of the American Society for Information Science and Technology 52(3), 226–234 (2001)
2. Sanderson, M., Kohler, J.: Analyzing geographic queries. In: Workshop on Geographic Information Retrieval SIGIR (2004)
3. Leidner, J.L.: Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names. Universal Press, Boca Raton, FL, USA (2008)
4. Andogah, G.: Geographically Constrained Information Retrieval. PhD thesis, University of Groningen, Groningen, Netherlands (2010)
5. Gruber T.R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer, The Netherlands, (1993)
6. Gruber T.R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2):199–220 (1993)
7. Haarslev V., Lutz C., Moller R.: Foundations of Spatioterminological Reasoning with Description Logics. Proceedings of the sixth Int. Conf. On Principles of Knowledge Representation and Reasoning (KR'98), A.G. Cohn et al, pp. 112-123 (1998)
8. Spaccapietra S., Cullot N., Parent C., Vangenot C.: On Spatial Ontologies. In: 6th Brazilalian Symposium on GeoInformatics, GeoInfo, Campos do Jordao, Brazil, Noviembre 22-24 (2004)
9. Tanasescu V., Gugliotta A., Domingue J., Davies R., Gutiérrez-Villarías L., Rowlatt M. Richardson M., Stincic S.: A Semantic Web Servicies GIS based Emergency Management

[5] http://www.racer-systems.com/products/porter/index.phtml

Application. International Semantic Web Conference, Athens, GA, USA, pp. 959-966, (2006)
10. Adriana Lopez, Maria J. Somodevilla, Darnes Vilarino, Ivo H. Pineda and Concepcion P. de Celis: Toponym Disambiguation by Ontology in Spanish: Geographical proximity between place names in the same context. In: AISS: Advances in Information Sciences and Service Sciences, Vol. 4, No. 1, pp. 282-289 (2012)
11. GeoWordNet: http://geowordnet.semanticmatching.org/
12. GeoNames: http://www.geonames.org/
13. Tripod: http://www.research-projects.uzh.ch/p8052.htm
14. Buscaldi D.: Toponym disambiguation in information retrieval. Procesamiento del Lenguaje Natural, [46]:125–126 (2011)
15. Budak Arpinar I., Amit Sheth, Cartic Ramakrishnan, E. Lynn Usery, Molly Azami & Mei-Po Kwan: Geospatial Ontology Development and Semantic Analytics. In: Handbook of Geographic Information Science, Eds: J. P. Wilson and A. S. Fotheringham, Blackwell Publishing (2004)
16. Hugo Gonçalo Oliveira, Diana Santos, Paulo Gomes & Nuno Seco: PAPEL: A lexical ontology for Portuguese. Springer-Verlag Berlin, Heidelberg (2008)
17. Fangling Jiang & Wenjun Wang: Research on Chinese Toponym Ontology model. In: 2010 International Conference of Information Science and Management Engineering, (2010)