Procesamiento del Lenguaje Natural, Revista nº 63, septiembre de 2019, pp. 143-146 recibido 03-04-2019 revisado 08-05-2019 aceptado 14-05-2019

Extracción automática de referencias geoespaciales en discurso libre usando técnicas de procesamiento de lenguaje natural y teoría de la accesibilidad

Extraction of Geospatial References from Free Text Based on Natural Language Processing and Accessibility Theory

Alejandro Molina-Villegas[1], **Oscar S. Siordia**[1],
Edwin Aldana-Bobadilla[2], **César Aguilar**[3], **Olga Acosta**[3]
[1]CONACyT – Centro de Investigación en Ciencias de Información Geoespacial, México
[2]CONACyT – Centro de Investigación y de Estudios Avanzados
del Instituto Politécnico Nacional, México
[3]Pontificia Universidad Católica de Chile, Chile
amolina@centrogeo.edu.mx

Resumen: La extracción automática de información geoespacial en tiempo real, a partir de discurso libre, tendrá un enorme impacto en tecnologías disruptivas, tales como los asistentes inteligentes y los motores de búsqueda. Generar modelos capaces de decodificar el discurso para luego transformarlo en datos estructurados aborda la solución de problemas como: la identificación automática de frases que se refieran a alguna entidad geoespacial; el manejo de sinónimos referidos a una misma locación (ambiguedad); la caracterización taxonómica de expresiones locativas; así como la automatización de procesos complejos de interpretación del lenguaje para la determinación de coordenadas geográficas en mapas y bases de datos. El proyecto presentado tiene por objetivo la automatización de procesos de georeferencia de documentos digitales para extraer conocimiento. La propuesta metodológica incluye compilación de un corpus, caracterización lingüística y desarrollo de algoritmos.
Palabras clave: Geoparsing, Geocodificación, Reconocimiento de Entidades Nombradas

Abstract: The automatic extraction of geospatial information in real time, from free speech, will an important impact on disruptive technologies, such as AI assistants and search engines. Models capable of decoding discourse and then transforming it into structured data addresses the solution of challenging problems such as: the automatic identification of phrases refering geospatial entities; the handling of synonyms referring to the same location (ambiguity); the taxonomic characterization of locative expressions; as well as the automation of complex processes of natural language interpretation to determine of geographical coordinates, maps and databases. The presented project aims to automate georeferencing processes of digital documents to extract georeferenced knowledge. The methodological proposal includes compilation of a corpus, linguistic characterization and algorithms.
Keywords: Geoparsing, Geocoding, Named Entity Recognition

1 Introducción

En los últimos años, el estudio de la expresión de nociones espaciales por medio del lenguaje natural, ha cobrado una gran relevancia, particularmente para la inteligencia artificial. Un buen ejemplo de esto son asistentes como Siri, Cortana o Alexa, por mencionar los más conocidos en el mercado. Uno de los motivos que subyace en tal relevancia —aunque no es el único— es la masificación de datos georreferenciados, implícita o explícitamente.

Este estudio propone la selección, recolección y procesamiento de expresiones en la variante de español mexicano (ya sean palabras o frases), que codifiquen algún concepto georreferencial asociable a entidades, escenarios o eventos locativos. En concreto, nos interesa

ISSN 1135-5948. DOI 10.26342/2019-63-16

resolver casos como:

1. *San Andrés Cholula es un municipio próximo a la Zona Metropolitana de Puebla.*
2. *El choque ocurrió esta mañana en el kilómetro 5 de la carretera México-Puebla*
3. *Los asaltantes irrumpieron en la sucursal que se encuentra en Puebla esquina con Monterrey, en la Colonia Roma*

En los tres casos, la entidad nominal *Puebla* tiene un claro atributo georreferencial, pero de diferente índole: mientras que (1) se refiere a un núcleo urbano, en (2) se alude a una vía que recorre dos provincias diferentes, mientras que (3) se refiere a dos calles situadas en un barrio de la Ciudad de México. Si bien es cierto que la ambigüedad que se genera para ubicar tal entidad en un mapa puede ser resuelta, también es verdad que resulta necesario delimitar el alcance de los atributos locativos asociados a dicha entidad, de tal suerte que puedan resolverse preguntas como: *¿cuál es la salida del metro más cercana a Puebla?*, en donde un asistente inteligente debería inferir que se trata de una calle, y no de una zona urbana, de la capital de una provincia, o de la provincia misma. En ese sentido, este proyecto aborda un problema *AI-Hard* que entrelaza la Inteligencia Artificial, la Lingüística y la Geomática, ya que en él se vislumbra la creación de métodos capaces de transformar voz o texto en identificadores geográficos inequívocos tales como latitud y longitud. Así, nuestra propuesta es multidisciplinaria, pues no solo se trata de caracterizar el fenómeno lingüístico de la georreferenciación discursiva en español, sino que también integrarlo en algoritmos y estructuras de datos de cara a la automatización completa de procesos sofisticados de georreferenciación por medio de la voz o el análisis de documentos de texto (*Geoparsing*). Con estas tecnologías lograríamos explotar, eficientemente, considerables cantidades de documentos existentes para extraer conocimiento georreferenciado.

2 Teoría de la accesibilidad

Un enfoque lingüístico pertinente para identificar entidades nominales con información georreferencial es el que plantea la *Teoría de la accesibilidad*, desarrollada principalmente por (2014), así como Gernsbacher y Givón (1995). Esta teoría explica cómo las unidades nominales y pronominales ofrecen un vínculo directo o indirecto a sus referentes, dependiendo de la cantidad de información que contengan. Tal cantidad de información, se sitúa en los niveles semántico y pragmático y permite que tales referentes sean reconocidos como elementos nuevos en el discurso (p. e., frases nominales largas como: *los vecinos de la nueva casa de enfrente*), o como elementos ya conocidos (p. e., el pronombre personal *ellos*, estableciendo una relación anafórica con la frase anterior). El tomar en cuenta la cantidad de información que contengan unidades nominales de índole locativa será útil para identificar aquellas que tengan un peso referencial relevante (p. e.: *la carretera federal México-Puebla*, con miras a contrastarlas con otras que puedan ser vistas o bien como segmentos nominales con valor anafórico situadas en un mismo contexto discursivo (*la México-Puebla*), o como unidades pronominales que requieren mayor información referencial para ser desambiguadas (p. e.: *por ahí se llega rápido a Puebla*).

3 Aportaciones del proyecto

Uno de los aportes relevantes de este proyecto será la caracterización lingüística de entidades georreferenciables. Para lo cual, nuestro enfoque metodológico considera la compilación del Corpus de Entidades Georreferenciadas de México (CEGEOMEX), el primero de su clase, el cual tendrá un anotado lingüístico manual, lo que facilitará el desarrollo de algoritmos de aprendizaje de máquina, así como la generación de meta-información discursiva. Cabe destacar que la anotación manual ya ha sido ampliamente utilizada en iniciativas internacionales tales como la CoNLL (*Conference on Natural Language Learning*) y que para este proyecto, resulta de especial interés la edición CoNLL 2002 (Sang y De Meulder, 2003) en donde por primera vez se consideró al español como una de las lenguas a procesar. Sin embargo, dado que en CoNLL 2002, los datos fueron recopilados por la Universidad Politécnica de Catalunya y la Universidad Autónoma de Barcelona, la anotación se focalizó en documentos de España, dejando de lado cualquier otra variante dialectal, entre ellas la mexicana. El rezago de recursos similares en México pone de manifiesto la necesidad de contar con un corpus de ca-

lidad con estas características. CEGEOMEX será un corpus de lengua general que se compondrá de documentos periodísticos, así como de segmentos de entrevistas y diálogos extraídos de la radio, con el propósito de conformar una colección balanceada de muestras escritas y orales siguiendo los criterios desarrollados por McEnery (2001) y Gries (2006) para observar variaciones en la expresión de entidades georreferenciadas en español mexicano. CEGEOMEX permitirá abordar la solución de problemas como: a) la identificación automática de frases u oraciones que se refieran a alguna entidad o evento de tipo espacial o georreferenciado implícitamente; b) el manejo de sinónimos referidos a una misma locación (Ciudad de México/Capital de la República/Distrito Federal/CDMX/DF); c) la propuesta de una taxonomía que ayude a clasificar expresiones locativas y entidades nombradas y d) la automatización de procesos complejos de interpretación del lenguaje para la determinación de coordenadas geográficas en mapas y bases de datos.

4 Objetivos

1. Establecer una caracterización lingüística de las entidades georreferenciadas mediante su definición formal, considerando tanto sus atributos lingüísticos, su estructuración en patrones, así como los parámetros contextuales que den indicios de su manifestación en un texto, sea oral o escrito.
2. Crear el primer Corpus de Entidades Georreferenciadas de México.
3. Innovar en la generación de algoritmos que combinando atributos lingüísticos y modelos computacionales detecten entidades georreferenciables para ser visualizadas en cartografía digital o en imágenes satelitales.
4. Consolidar un grupo de investigación multidisciplinario e internacional para desarrollar proyectos relacionados con nuevas áreas de investigación que unifiquen computación, lingüística y geomática.
5. Formar recursos humanos especializados en investigación de frontera.

5 Metas

1. Crear un repositorio de archivos para almacenar documentos periodísticos de México que incluya, al menos cinco de los periódicos principales de cobertura nacional: *El Universal*, *La Jornada*, *El Financiero*, *El Sol de México*, *La Razón*, *Uno Más Uno* y *Reforma*; así como seis estaciones de radio: *MVS noticias*, *Radio Fórmula*, *Radio IPN*, *Radio UNAM*, *Radio Ibero*, *Red FM* y Stereo Cien.
2. Implementar el primer Corpus de Entidades Georreferenciadas de México (CEGEOMEX) que servirá de base para futuros proyectos.
3. Diseñar un sistema para detectar entidades Georreferenciadas en discurso libre (oral y escrito).

6 Metodología

La metodología propuesta cubre 3 etapas, las cuales son descritas a continuación.

6.1 Etapa I

En aras de compilar un volumen masivo de documentos para caracterizar el fenómeno de estudio, en sus variantes escrita y oral, se desarrollarán programas informáticos para automatizar procesos de descarga, almacenamiento, indexado, transcripción y organización de noticias en los medios de comunicación antes mencionados, empleando técnicas de *Web Crawling*. Dado que se procesarán documentos de texto y audio, será necesario adquirir y configurar equipo especializado y software para ejecutar tal procesamiento, de tal suerte que se evaluarán e incluirán modelos de reconocimiento de voz para la transcripción de noticias de radio en español mexicano. El equipo encargado del análisis lingüístico estará integrado por investigadores de la Pontificia Universidad Católica de Chile, quienes han abordado el análisis de relaciones espaciales reconocidas entre términos médicos en español, tomando en cuenta un enfoque cognitivo (ver Acosta y Aguilar (2015)). Los resultados de su análisis refuerzan la hipótesis de que la concepción espacial se extiende a muchos dominios abstractos, tales como tiempo, estado, posesión, corporeidad u organización social, entre los más relevantes. La tarea de estos investigadores será aportar al proyecto una propuesta de taxonomía y especificación de rasgos espaciales para el diseño del etiquetado de entidades georreferenciadas. Estos trabajos ayudarán a

establecer el metalenguaje (etiquetas y atributos) y las consideraciones para el anotado manual del corpus.

6.2 Etapa II

En esta etapa se creará el corpus CEGEOMEX, el cual se etiquetará manualmente. Para apoyar esta tarea, se desarrollará una plataforma de anotación en el Centro de Investigación en Ciencias de Información Geoespacial (México). Dicha herramienta brindará la especificación de las etiquetas y atributos obtenidos de la etapa anterior. La herramienta dará acceso remoto y multiusuario de manera que en seis meses se llegue a la meta de, al menos, 12000 colocaciones anotadas, divididas en 6000 instancias orales y 6000 instancias escritas. Cabe mencionar que ningún corpus anotado cuenta actualmentelas características específicas de esta investigación.

6.3 Etapa III

Finalmente, en esta etapa se desarrollarán los algoritmos y un software que concentrará los resultados de la investigación (Figura 1). Aprovechando los resultados obtenidos de las exploraciones que se hagan al CEGEOMEX, será posible experimentar con algoritmos híbridos que utilizarán tanto atributos simbólicos (tales como posición de entidades y categorías gramaticales), así como variables abstractas generadas mediante técnicas de aprendizaje de máquina y reconocimiento de patrones. El hecho de generar variables abstractas será útil para vincular y representar en forma de vectores los resultados obtenidos (p. e., candidatos a entidades georreferenciales codificadas en nombres o en frases nominales). Dichos vectores, además de ofrecer una descripción numérica sobre el comportamiento de los resultados obtenidos, serán valiosos para categorizar similitudes semánticas identificables entre tales unidades lingüísticas. Cabe señalar aquí que estudios previos se ha demostrado la eficiencia de combinar patrones lingüísticos con algoritmos de aprendizaje de máquina (Sierra et al., 2009).

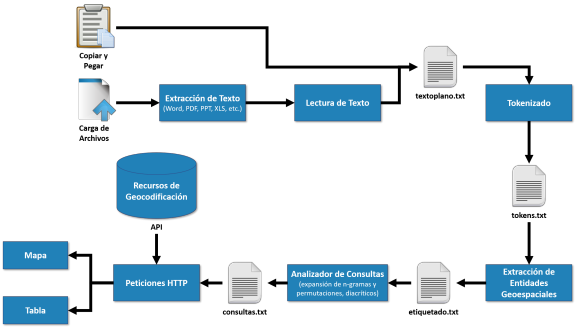

Figura 1: Proceso de extracción de referencias geoespaciales en discurso libre. Los documentos son preprocesados y tokenizados para que un módulo de Reconocimiento de Entidades Nombradas detecte nombres de lugares que serán georreferenciados mediante recursos externos para obtener sus coordenadas

Agradecimientos

Proyecto FORDECyT 296737 (Consorcio en Inteligencia Artificial) y a la Red Temática en Tecnologías del Lenguaje por el financiamiento parcial de esta investigación. A la Mtra. en Literatura Mariana Tello-Signoret por la revisión de datos.

Bibliografía

Acosta, O. y C. A. Aguilar. 2015. Extracting concrete entities through spatial relations. En *Proceedings of the 3rd International Workshop on Artificial Intelligence and Cognition, Turin, Italy, September 28-29, 2015.*, páginas 133–145.

Ariel, M. 2014. *Accessing noun-phrase antecedents*. Routledge.

Gernsbacher, M. A. y T. Givón. 1995. *Coherence in spontaneous text*, volumen 31. John Benjamins Publishing.

Gries, S. T. 2006. Exploring variability within and between corpora: some methodological considerations. *Corpora*, 1(2):109–151.

McEnery, T. 2001. Corpus linguistics/tony mcenery, andrew wilson.

Sang, E. F. y F. De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. *arXiv preprint cs/0306050*.

Sierra, G.-E., R. Alarcon, A. Molina-Villegas, y E. Aldana. 2009. Web exploitation for definition extraction. En *IEEE Latin American Web Congress, doi 10.1109/LA-WEB.2009.36*, páginas 217–223.