Procesamiento del Lenguaje Natural, núm. 43 (2009), pp. 33-40 recibido 1-05-2009; aceptado 5-06-2009

# Geo-NER: un reconocedor de entidades geográficas para inglés basado en GeoNames y Wikipedia*

## *Geo-NER: an English geographic entity recognizer based on GeoNames and Wikipedia*

**José Manuel Perea Ortega** **Arturo Montejo Ráez**
**Fernando Martínez Santiago** **L. Alfonso Ureña López**

Departamento de Informática, Escuela Politécnica Superior
Universidad de Jaén, E-23071 - Jaén
{jmperea,amontejo,dofer,laurena}@ujaen.es

**Resumen:** En este artículo se presenta una herramienta para detectar y reconocer entidades específicamente geográficas. Esta herramienta se basa en la utilización de recursos externos tales como *gazetteers*, Wikipedia y reconocedores de entidades genéricos. Durante su funcionamiento se hace uso de cierto razonamiento geográfico, apoyándose en patrones sintácticos y características geográficas (términos tales como *río*, *montaña*, *lago*, etc.). Para su evaluación se han utilizado las consultas geográficas desarrolladas para la tarea GeoCLEF, enmarcada en las conferencias CLEF, obteniendo prometedores resultados para el reconocimiento de entidades de tipo geográfico, en comparación con otros reconocedores de entidades genéricos.
**Palabras clave:** Reconocimiento de entidades nombradas, recuperación de información geográfica

**Abstract:** In this paper we show a tool for detecting and recognizing geographic entities specifically. This tool is based on the use of external resources such as gazetteers, Wikipedia and generic entities recognizers. During its operation, it makes use of geographical reasoning, based on syntactic patterns and geographic features (terms such as *river*, *mountain*, *lake*, etc.). The evaluation was carried on the geographic queries developed for the task GeoCLEF, part of the CLEF conferences, obtaining promising results in the recognition of geographic entities, compared to other generic entity recognizers.
**Keywords:** Named entities recognition, geographic information retrieval

## *1. Introducción*

Hoy día, la información geográfica se encuentra presente en una amplia variedad de medios y tipos de documentos. Durante las pasadas décadas, la tecnología empleada para acceder a este tipo de información se ha centrado en la combinación de mapas digitales y bases de datos, caracterizando a la mayoría de los Sistemas de Información Geográfica (*Geographic Information Systems*, GIS) (Chang, 2007; Bolstad, 2005). Sin embargo, en los últimos años se ha prestado especial atención al desarrollo de sistemas automáticos que recuperen información específicamente geográfica presente en documentos de texto no estructurados como los que componen la Web (Larson, 1996; McCurley, 2001; Jones y Purves, 2008).

La recuperación de información geográfica (*Geographic Information Retrieval*, GIR) puede considerarse una extensión de la recuperación de información tradicional (*Information Retrieval*, IR), incluyendo todas las áreas que tradicionalmente forman el núcleo de investigación en IR y haciendo un especial énfasis en el indexado y recuperación espacial y geográfica (Larson, 1996). Se ha comprobado que una de las partes fundamentales en una arquitectura GIR es el motor de recuperación de información utilizado (Perea-Ortega et al., 2008a). Un análisis general sobre los sistemas GIR presentados a la tarea

* Esta investigación ha sido parcialmente financiada por el Gobierno Español, proyecto TIMMOM (TIN2006-15265-C06-03), por la Universidad de Jaén, proyecto RFC/PP208/UJA-08-16-14 y por la Junta de Andalucía, Consejería de Turismo y Deporte (FFIEXP06-TU2301-2007/000024)

 

GeoCLEF durante los años 2005 a 2007 ha sido realizado en Perea-Ortega et al. (2008b). En definitiva, la tarea GIR se centra en la mejora de la calidad de la recuperación de información específicamente geográfica para el acceso a documentos no estructurados (Jones y Purves, 2008).

Otro aspecto fundamental en un sistema GIR es la detección de las entidades geográficas presentes tanto en el texto de la colección como en la consulta de usuario. Detectar referencias a items geográficos permite aplicar una amplia gama de técnicas y herramientas. A modo de ejemplo, algunas de las técnicas usuales en la literatura y que requieren de un buen reconocedor de entidades geográficas son: añadir una signatura a cada documento con su ámbito geográfico (Martins y Silva, 2005; Silva et al., 2006), crear índices invertidos geográficos (Martins, Silva, y Andrade, 2005; Li et al., 2006) o reformular la consulta de usuario en dos, una genérica y otra geográfica.

Por otra parte, existen gran cantidad de reconocedores de entidades genéricos, pero éstos se muestran limitados cuando se trata de reconocer entidades geográficas ambiguas, o referencias indirectas (por ejemplo, un gentilicio) o cuando refieren áreas geográficas vagas ("*Western Europe*", por ejemplo).

El artículo se organiza de la siguiente manera: en primer lugar, se explican algunos conceptos y el trabajo relacionado. Seguidamente, se describe la arquitectura y el funcionamiento básico de la herramienta presentada en este trabajo. A continuación, se muestran los experimentos y resultados obtenidos, haciendo un análisis comparativo. Finalmente, se comentan las conclusiones y el trabajo futuro.

## 2. Conceptos y trabajo relacionado

Entendemos por *geo-entidad* todo término que hace referencia a un nombre de lugar, o lo que es lo mismo, un topónimo, un lugar geográfico. Determinar qué consideramos como *geo-entidad* o referencia geográfica en cualquier texto puede resultar, a veces, una tarea bastante compleja (Santos y Chaves, 2006). La habilidad para reconocer y razonar sobre la terminología geográfica presente en cualquier texto es un aspecto crucial en un sistema GIR (Jones et al., 2002).

La tarea de extraer y distinguir diferentes tipos de entidades en documentos de texto se conoce normalmente como Reconocimiento de Entidades Nombradas (*Named Entity Recognition*, NER). En las últimas décadas ha sido una tarea de minería de texto bastante importante, en la que se ha conseguido un funcionamiento automático satisfactorio. Sin embargo, el problema específico de reconocer referencias geográficas presenta desafíos adicionales (Kornai y Sundheim, 2003). Cuando manejamos nombres de entidades con un alto nivel de detalle, los problemas de ambigüedad surgen con más frecuencia. Existen dos tipos de ambigüedad para los nombres de lugar (Amitay et al., 2004): cuando la geo-referencia tiene un significado no geográfico (ej. *Reading* es una ciudad de Inglaterra), y otra, cuando dos referencias geográficas poseen el mismo nombre (ej. en EEUU existen 18 ciudades que se llaman *Jerusalem*). Algunos trabajos previos se han centrado en etiquetar las localizaciones presentes en páginas Web, así como en asignar ámbitos geográficos a las mismas (Silva et al., 2006; Martins y Silva, 2005; Jones et al., 2002). Otros, han aplicado técnicas de clasificación automática y minería de texto a consultas lanzadas a diferentes motores de búsqueda (Li, Zheng, y Dai, 2005; Vogel et al., 2005; Andogah et al., 2008).

Para la herramienta presentada en este trabajo, el concepto de *geo-entidad* intenta abarcar algo más que el reconocimiento de nombres de lugar. Por ejemplo, ¿deberíamos considerar como *geo-entidad* un simple gentilicio presente en una sentencia? Si analizamos la sentencia "*Ship traffic around the Portuguese islands*" podemos observar que realmente no aparece ningún nombre de lugar, pero sin embargo, si nuestro detector de geo-entidades fuera capaz de reconocer *Portuguese* como un gentilicio y, además, comprobar que el término que le sigue es una característica geográfica (isla), esto sí nos aportaría cierto conocimiento geográfico sobre la consulta (el usuario está interesado en islas de Portugal). Sin embargo, si analizamos la sentencia "*American troops in the Persian Gulf*", observamos que aparece el gentilicio "*American*" pero el término que le sigue a continuación no es ninguna característica geográfica (*tropas*) y por tanto no debería considerarse una *geo-entidad*.

Por otro lado, existen casos en los que la referencia geográfica no aparece bien delimitada. Por ejemplo, en la sentencia "*Natural*

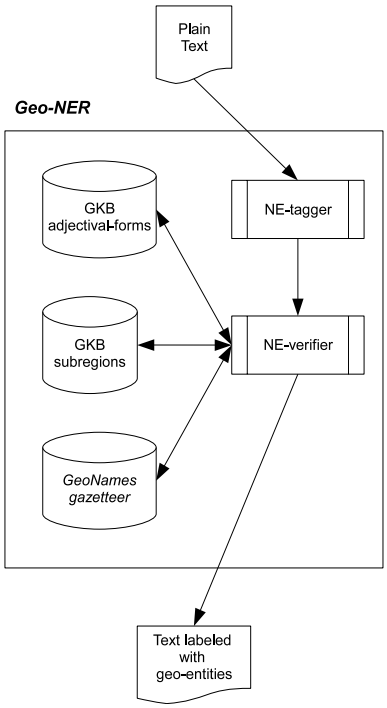


Figura 1: Arquitectura de Geo-NER

*disasters in the Western USA*", ¿deberíamos considerar como geo-entidad *USA* o *Western USA*?. Si nuestra herramienta es capaz, además, de detectar posibles subregiones del mundo que ya están definidas y reconocidas por la División Estadística de las Naciones Unidas[1], la precisión en la detección y reconocimiento de las referencias geográficas sería más alta. Además, se reduciría la ambigüedad existente cuando no conocemos si el ámbito geográfico de un documento pertenece a la geo-entidad detectada en la consulta.

Debido a los tipos de ambigüedad presentes en algunas geo-referencias, como se ha comentado con anterioridad, si en la tarea GIR optásemos por utilizar un NER genérico para el reconocimiento de *geo-entidades*, se podría dar el caso de que ciertas localizaciones fueran detectadas como organizaciones o incluso como nombres de personas. Para evitar esto, sería conveniente que nuestro detector de geo-entidades fuera capaz de comprobar aquellas entidades no reconocidas como localizaciones con el fin de aumentar no sólo la precisión, sino también la cobertura.

[1] http://unstats.un.org/unsd/

## 3. *Geo-NER: arquitectura y funcionamiento*

En pocas palabras, el reconocedor de entidades geográficas Geo-NER toma como base un etiquetador de entidades nombradas tradicional, el cual es enriquecido mediante el uso de varias fuentes de conocimiento y heurísticas específicas del ámbito de la terminología geográfica. A continuación, se detalla la arquitectura de esta herramienta, que se compone de varios elementos, como se muestra en la Figura 1:

- **Etiquetador de entidades (*NE-tagger*)**. Utilizamos el *etiquetador de entidades nombradas LBJ* creado por el grupo *Cognitive Computation* de la Universidad de Illinois (Li, Morie, y Roth, 2005). Hace uso de varios tipos de etiqueta: persona, organización, localización y miscelánea. Usa diferentes *gazetteers* extraídos de la Wikipedia[2], clases de palabras derivadas de textos no etiquetados y características expresivas no locales. Su mejor resultado es un 90.1 de medida F1 sobre los datos de la tarea compartida del CoNLL03[3]. Es un etiquetador robusto que ha sido evaluado sobre diferentes conjuntos de datos.
- **Base de conocimiento geográfico sobre gentilicios (*GKB-adjectival-forms*)**. Es una base de datos formada por gentilicios y sus países correspondientes. El proceso de construcción de este recurso se explica en la siguiente sección.
- **Base de conocimiento geográfico sobre subregiones (*GKB-subregions*)**. Es otra base de datos generada a partir de las subregiones del mundo definidas por Naciones Unidas. En la siguiente sección se explica cómo ha sido generada esta base de conocimiento.
- ***Gazetteer***. Utilizamos el recurso externo *GeoNames*[4] como base de datos de nombres de lugar. Contiene cerca de 8 millones de nombres geográficos, incluyendo 2.2 millones de lugares populares y 1.8 millones de nombres alternativos.

[2] http://www.wikipedia.org/
[3] http://www.cnts.ua.ac.be/conll2003/
[4] http://www.geonames.org/

- **Verificador de entidades (*NE-verifier*)**. Es el proceso encargado de comprobar, para todas las entidades etiquetadas por el *NE-tagger*, si realmente son geo-entidades. Utiliza algunas heurísticas como emparejamiento simple, patrones sintácticos y características geográficas, apoyándose en las bases de conocimiento geográfico comentadas anteriormente.

### 3.1. Base de conocimiento geográfico

Para construir la base de conocimiento geográfico en la que se basa Geo-NER se han utilizado técnicas de extracción de información en un recurso externo como la Wikipedia. Para el tratamiento de gentilicios hacemos uso de la lista de los adjetivos de nombres de lugar presente en Wikipedia[5]. Para el tratamiento de las subregiones del mundo hacemos uso de la definición de regiones continentales y subregiones dada por Naciones Unidas. Estas subregiones se han definido para propósitos estadísticos y se pueden consultar en Wikipedia[6].

Por otro lado, para la utilización del gazetteer GeoNames hacemos uso de los diferentes servicios Web[7] que ofrece, permitiéndonos realizar consultas óptimas a su base de datos. GeoNames categoriza todas las características geográficas posibles entre una de las nueve clases principales:

- A para países, estados, regiones.
- H para ríos, lagos, mares...
- L para parques, áreas...
- P para ciudades, pueblos, villas...
- R para carreteras
- S para edificios, granjas...
- T para montañas, colinas, rocas...
- U para zonas submarinas...
- V para zonas boscosas

Además, utiliza 645 subcategorías o códigos de categoría dentro de estas 9 clases principales. Algunos ejemplos de estas subclases son: *PPLC* para capitales de un entidad política, *LK* para lagos, *SEA* para mares, *DSRT* para desiertos, etc. Para los experimentos llevados a cabo en este artículo, sólo se han utilizado referencias geográficas que pertenecen a cinco de las nueve clases principales (A, P, L, H y T). El resto de clases no se han tomado en consideración para así facilitar la desambigüación de las entidades, basándonos en trabajos previos como (Andogah et al., 2008; Rauch, Bukatin, y Baker, 2003).

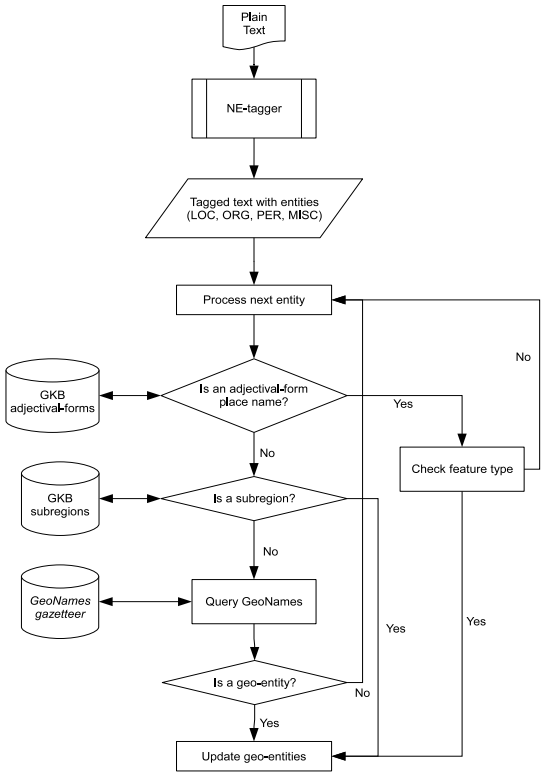


Figura 2: Geo-NER: algoritmo de funcionamiento

### 3.2. Funcionamiento de Geo-NER

El algoritmo que describe el funcionamiento básico de Geo-NER se puede observar en la Figura 2. Una vez que el etiquetador de entidades reconoce las entidades del texto de entrada, vamos procesando las entidades de todos los tipos mediante los siguientes pasos:

1. Comprobamos si se trata de un gentilicio. Este proceso consiste en buscar en la base de datos de gentilicios la entidad a procesar. Si se encuentra, se comprueba mediante emparejamiento simple si los

[5] http://en.wikipedia.org/wiki/List_of_adjectival_forms_of_place_names

[6] http://en.wikipedia.org/wiki/United_Nations_geoscheme

[7] http://www.geonames.org/export/ws-overview.html

términos que vienen después del gentilicio detectado se corresponden con alguna de las más de 700 características geográficas reconocidas por GeoNames (*city*, *country*, *mountain*, *islands*, etc.). Si coincide, se trata de una *geo-entidad*. De lo contrario, se procesa la siguiente.

2. Si la entidad a procesar no es un gentilicio, comprobamos si se trata de una subregión, buscándola en la base de conocimiento geográfico de subregiones. Si es una subregión, se acepta como *geo-entidad*.

3. En el caso de que no fuera subregión, se consulta a GeoNames para comprobar si se trata realmente de una localización. Si la consulta a GeoNames encuentra dicha entidad, aplicamos un enfoque mediante heurísticas basadas en la importancia del lugar y en su tipo geográfico. Por defecto, la desambiguación de una entidad geográfica puede ser determinada siguiendo varios enfoques: según la ocurrencia de lugar más común (Smith y Mann, 2003), por la población de la localización (Rauch, Bukatin, y Baker, 2003) o por extracción semi-automática desde la Web (Li et al., 2003). En el sistema se ha optado por la segunda heurística, basada en la población del lugar. Es decir, se acepta una entidad como geográfica cuando su clase principal (tipo) es A o P y GeoNames tiene almacenado el dato de su población, siendo ésta mayor que 0.

   El resto de heurísticas utilizadas en la herramienta son:

   - Se acepta la entidad como geográfica si su clase principal es *H* (corriente de agua) y su subcategoría o código de categoría comienza por "*ST*" (río) o es directamente *OCN* (océano), *SEA* (mar) o *GULF* (golfo). Por ejemplo, para la sentencia "*Shipwrecks in the Atlantic Ocean*", al consultar GeoNames con la entidad detectada por el *NE-tagger* (*Atlantic Ocean*), éste nos devuelve un registro de clase *H* y subclase *OCN*, con lo que la entidad es etiquetada como geográfica.
   - Si el tipo de la entidad es *T* y su subclase comienza por *MT* (sistema montañoso) o es *DSRT* (desierto), se acepta dicha entidad como localización. Por ejemplo, para la consulta "*Russian troops in the southern Caucasus*", *Caucasus* es encontrado en GeoNames como una entidad de clase *T* y subclase *MTS*.
   - Si el tipo de la entidad es *L* y su subclase es *CONT* (continente), también se acepta la entidad como geográfica. Para la sentencia "*Snowstorms in North America*" tenemos un ejemplo en el que la entidad *North America* aparece en GeoNames como de tipo *L* y subclase *CONT*.

Es conveniente aclarar que la herramienta propuesta es capaz de reconocer entidades que se encuentran en alguna de las fuentes de conocimiento mencionadas, pero no tiene capacidad de generalización, es decir, sólo puede detectar entidades conocidas de antemano y almacenadas en las bases de conocimiento.

## 4. Experimentos y resultados

Para los experimentos llevados a cabo en este artículo se han utilizado las consultas propuestas en la competición GeoCLEF durante los años de 2005 a 2008 (Gey et al., 2005; Gey et al., 2006; Mandl et al., 2007). El objetivo de este foro de evaluación es proporcionar un marco para evaluar sistemas GIR. En total se ha trabajado con 100 consultas textuales (25 por cada año). Estos topics son descripciones o requerimientos de usuario con tres campos principales: título (*title*), descripción (*desc*) y narrativa (*narr*). Para los experimentos se han utilizado únicamente los dos primeros campos (título y descripción). Un ejemplo de consulta se muestra en la Figura 3.

En primer lugar, se ha procedido a etiquetar manualmente todas las consultas, detectando las entidades geográficas en cada una de ellas. Posteriormente, se han ejecutado varios reconocedores de entidades tales como GATE[8], LingPipe[9] y el etiquetador de entidades LBJ del grupo *Cognitive Computation* de la Universidad de Illinois[10]. Finalmente, se han obtenido los resultados uti-

[8] http://gate.ac.uk/

[9] http://alias-i.com/lingpipe/

[10] http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=FLBJNE

```
<top>
 <num>10.2452/58-GC</num>
    <title>Travel problems at major
     airports near to London</title>
    <desc>To be relevant, documents
     must describe travel problems
     at one of the major airports
     close to London.</desc>
    <narr>Major airports to be listed
     include Heathrow, Gatwick, Luton,
     Stanstead and London City
     airport.</narr>
</top>
```

Figura 3: Ejemplo de una consulta GeoCLEF

lizando la herramienta Geo-NER descrita en este artículo. Además, para los experimentos realizados con esta herramienta, se ha comprobado qué mejora aporta el uso de cada uno de los recursos geográficos generados, tanto con la base de conocimiento geográfico sobre gentilicios (*GKB-adjectival-forms*) como con la de subregiones (*GKB-subregions*), así como la mejora aportada por el uso de las heurísticas definidas.

En relación a los diferentes reconocedores de entidades genéricos empleados, se han utilizado las configuraciones por defecto de los mismos. Para GATE se ha trabajado con la versión 4.0, haciendo uso del sistema de extracción de información *ANNIE* que contiene esta versión. Se han anotado las consultas utilizando su *tokenizador*, filtrando por entidades de tipo persona, localización y organización. Además, se ha utilizado el *gazetteer* incorporado en GATE. Para LingPipe se ha trabajado con la versión 3.4.0. Este detector de entidades nombradas está basado en reglas y expresiones regulares.

La Tabla 1 muestra estos experimentos y los resultados obtenidos en valores de precisión (P), cobertura (*recall*, R) y medida F, donde F es la media armónica de la precisión y la cobertura:

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}} \quad (1)$$

Analizando los resultados, si comparamos los obtenidos por Geo-NER con el resto de reconocedores empleados, se puede observar una mejora significativa en cuanto a precisión y cobertura al usar esta herramienta. En valores de medida F, se obtienen importantes

| NER | Recall | Prec. | F |
|---|---|---|---|
| LingPipe | 0.64 | 0.76 | 0.69 |
| GATE | 0.64 | 0.88 | 0.74 |
| LBJ-NE-Tagger | 0.74 | 0.92 | 0.82 |
| Geo-NER sin subregiones | 0.87 | 0.92 | 0.89 |
| Geo-NER sin gentilicios | 0.83 | 0.97 | 0.90 |
| Geo-NER sin heuristicas | 0.80 | 0.92 | 0.86 |
| Geo-NER | 0.97 | 0.98 | **0.97** |

Tabla 1: Experimentos y resultados

mejoras en torno al 40 %, 31 % y 18 % con respecto al uso del LingPipe, GATE y LBJ-NE-Tagger respectivamente. La precisión y la cobertura usando Geo-NER es superior al resto de reconocedores, obteniendo una diferencia mayor con respecto a la cobertura (40 % mejor que LingPipe y GATE y 31 % mejor que LBJ-NE-Tagger). Esto es debido a la utilización de todos los tipos de entidades (no sólo de las de tipo localización) durante el proceso de verificación (*NE-verifier*).

Evaluando lo que aportan las fuentes de conocimiento geográfico generadas, podemos observar que:

- Respecto al mejor NER genérico (LBJ-NE-Tagger), Geo-NER sin *GKB-adjectival-forms* (gentilicios) lo supera en un 9.7 %. Por otro lado, el uso de Geo-NER sin *GKB-subregions* (recurso sobre subregiones del mundo), también mejora en un 8.5 % la medida F del LBJ-NE-Tagger. Por último, Geo-NER sin aplicar las heurísticas definidas también lo mejora en un 4.8 %.
- Respecto al propio Geo-NER, se aprecia que el utilizar dichos recursos y heurísticas mejora el comportamiento general del sistema en torno al 12.8 %, en valor de medida F.

En definitiva, Geo-NER es el único de todos los sistemas evaluados que supera el 90 % en términos de precisión, cobertura y F-medida, con una mejora media de 0.15 puntos (lo que supone un 18 % más) respecto al mejor de los NER evaluados. El efecto de mejora se produce, principalmente, sobre la cobertura, aunque sin dañar la precisión, puesto que también se mejora.

## 5. Conclusiones y trabajo futuro

En este artículo se presenta una herramienta para la detección y reconocimiento de entidades de tipo geográfico, basada en un etiquetador de entidades genérico y en diversos recursos de conocimiento geográfico que también han sido generados utilizando la Wikipedia. Esta herramienta también tiene como base de funcionamiento un *gazetteer* como GeoNames, apoyándose en diversas heurísticas desarrolladas. Para demostrar su efectividad, se han utilizado las 100 consultas generadas para el foro de evaluación GeoCLEF, donde aparecen entidades de diferentes tipos, no sólo geográficas. Mediante su utilización, se consigue mejorar tanto la cobertura como la precisión de varios reconocedores de entidades genéricos como GATE o LingPipe. La cobertura se mejora debido a que Geo-NER utiliza todas las etiquetas del reconocedor base con el que trabaja, haciendo que aquellas entidades que no han sido marcadas como localizaciones puedan serlo si cumplen alguna de las heurísticas preestablecidas. Por otro lado, la precisión también se mejora como se ha demostrado, debido al uso de los recursos geográficos generados mediante Wikipedia y al uso de un *gazetteer* bastante completo como GeoNames, permitiendo comprobar si la entidad a procesar es realmente una entidad geográfica.

Como trabajo futuro se pretende ampliar los recursos geográficos generados para contemplar más subregiones definidas en el mundo, evaluar la herramienta con otras colecciones o corpus, así como probar un detector de entidades basado en n-gramas en lugar de utilizar un etiquetador genérico. El uso de información sintáctica para la detección de posibles entidades debería ser investigado. También se pretende estudiar la aplicación de Geo-NER a partir de otros reconocedores génericos como GATE o LingPipe.

## Bibliografía


Amitay, Einat, Nadav Har'El, Ron Sivan, y Aya Soffer. 2004. Web-a-where: Geotagging web content. En Mark Sanderson Kalervo Järvelin James Allan, y Peter Bruza, editores, *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 273–280, Sheffield, UK, July. ACM Press.

Andogah, Jeffrey, Gosse Bouma, John Nerbonne, y Erwin Koster. 2008. Place-name ambiguity resolution. En *LREC 2008 workshop on Methodologies and Resources for Processing Spatial Language*.

Bolstad, P. 2005. *GIS Fundamentals: A first text on Geographic Information Systems, Second Edition*. Eider Press.

Chang, K. 2007. *Introduction to Geographic Information System, 4th Edition*. McGraw-Hill College.

Gey, Fredric, Ray Larson, Mark Sanderson, Hideo Joho, Paul Clough, y Vivien Petras. 2005. Geoclef: the clef 2005 cross-language geographic information retrieval track overview. En *Sixth Workshop of the Cross-Language Evaluation Forum: Working Notes for the CLEF 2005 Workshop (CLEF'2005)*, página s/ pp., Viena and Austria and 21-23 de Setembro de 2005.

Gey, Fredric C., Ray Larson, Mark Sanderson, Kerstin Bischoff, Thomas Mandl, Christa Womser-Hacker, Diana Santos, Paulo Rocha, Giorgio Maria Di Nunzio, y Nicola Ferro. 2006. Geoclef 2006: The clef 2006 cross-language geographic information retrieval track overview. En Carol Peters Paul Clough Fredric C. Gey Jussi Karlgren Bernardo Magnini Douglas W. Oard Maarten de Rijke, y Maximilian Stempfhuber, editores, *CLEF*, volumen 4730 de *Lecture Notes in Computer Science*, páginas 852–876. Springer.

Jones, Christopher B., Ross Purves, Anne Ruas, Mark Sanderson, Monika Sester, Marc J. van Kreveld, y Robert Weibel. 2002. Spatial information retrieval and geographical ontologies an overview of the spirit project. En *SIGIR*, páginas 387–388. ACM.

Jones, Christopher B. y Ross S. Purves. 2008. Geographical information retrieval. *International Journal of Geographical Information Science*, 22(3):219–228.

Kornai, Andras y Beth Sundheim, editores. 2003. *Workshop on the Analysis of Geographic References*. (held in conjunction with NAACL-HLT 2003).

Larson, R. 1996. Geographic information retrieval and spatial browsing. En Smith y M. Gluck, editores, *Geographic Information Systems and Libraries: Patrons*

*nd Maps and Spatial Information*, páginas 81–124.

Li, H., K. R. Srihari, C. Niu, y W. Li. 2003. Infoxtract location normalization: a hybrid approach to geographic references in information extraction. En A. Kornai y B. Sundheim, editores, *HLT-NAACL 2003 Workshop Analysis of Geographic References*, páginas 39–44, Edmontonand Albertaand Canada, May 31. Association for Computational Linguistics.

Li, X., P. Morie, y D. Roth. 2005. Semantic integration in text: From ambiguous names to identifiable entities. *AI Magazine. Special Issue on Semantic Integration*, páginas 45–68.

Li, Ying, Zijian Zheng, y Honghua (Kathy) Dai. 2005. KDD CUP-2005 report: Facing a great challenge. *SIGKDD Explorations*, 7(2):91–99.

Li, Zhisheng, Chong Wang, Xing Xie, Xufa Wang, y Wei-Ying Ma. 2006. Indexing implicit locations for geographical information retrieval. En Ross Purves y Chris Jones, editores, *GIR*. Department of Geography, University of Zurich.

Mandl, Thomas, Fredric C. Gey, Giorgio Maria Di Nunzio, Nicola Ferro, Ray Larson, Mark Sanderson, Diana Santos, Christa Womser-Hacker, y Xing Xie. 2007. Geoclef 2007: The clef 2007 cross-language geographic information retrieval track overview. En Carol Peters Valentin Jijkoun Thomas Mandl Henning Müller Douglas W. Oard Anselmo Peñas Vivien Petras, y Diana Santos, editores, *CLEF*, volumen 5152 de *Lecture Notes in Computer Science*, páginas 745–772. Springer.

Martins, Bruno, Mario J. Silva, y Leonardo Andrade. 2005. Indexing and ranking in geo-ir systems. En *Proc. of the workshop on geographic information retrieval - GIR'05*, páginas 31–34, New York, NY, USA. ACM Press.

Martins, Bruno y Mário J. Silva. 2005. A graph-ranking algorithm for geo-referencing documents. En *ICDM*, páginas 741–744. IEEE Computer Society.

McCurley, K. S. 2001. Geospatial mapping and navigation of the web. En *Proc. of the Tenth International World Wide Web*, páginas 221–229, Hong Kong, May 1-5.

Perea-Ortega, José M., Miguel Angel García Cumbreras, Manuel García Vega, y L.Alfonso Ureña López. 2008a. Comparing several textual information retrieval systems for the geographical information retrieval task. En Epaminondas Kapetanios Vijayan Sugumaran, y Myra Spiliopoulou, editores, *NLDB*, volumen 5039 de *Lecture Notes in Computer Science*, páginas 142–147. Springer.

Perea-Ortega, José M., Miguel Angel García Cumbreras, Manuel García Vega, y L.Alfonso Ureña López. 2008b. Sistemas de recuperación de información geográfica multilingües en clef. *Sociedad Española para el Procesamiento del Lenguaje Natural*, (40):129–136.

Rauch, E., M. Bukatin, y K. Baker. 2003. A confidence-based framework for disambiguating geographic terms. En A. Kornai y B. Sundheim, editores, *HLT-NAACL 2003 Workshop Analysis of Geographic References*, Edmontonand Albertaand Canada. Association for Computational Linguistics.

Santos, Diana y Marcirio Silveira Chaves. 2006. The place of place in geographical ir. En Ross Purves y Chris Jones, editores, *GIR*. Department of Geography, University of Zurich.

Silva, M.J., B. Martins, M. Chaves, N. Cardoso, y A.P. Afonso. 2006. Adding geographic scopes to web resources. *CEUS-Computers, Environment and Urban Systems*, 30(378-399):93.

Smith, David A. y Gideon S. Mann. 2003. Bootstrapping toponym classifiers. En *Proceedings of the HLT-NAACL 2003 workshop on Analysis of geographic references*, páginas 45–49, Morristown, NJ, USA. Association for Computational Linguistics.

Vogel, David S., Steffen Bickel, Peter Haider, Rolf Schimpfky, Peter Siemen, Steve Bridges, y Tobias Scheffer. 2005. Classifying search engine queries using the web as background knowledge. *SIGKDD Explorations*, 7(2):117–122.