Procesamiento del Lenguaje Natural, Revista nº 40, marzo de 2008, pp. 129-136 recibido 08-02-08, aceptado 03-03-08

Sistemas de Recuperación de Información Geográfica multilingües en CLEF*

Multilingual Geographical Information Retrieval systems in CLEF

José Manuel Perea Ortega **Miguel Angel García Cumbreras**
Manuel García Vega **L. Alfonso Ureña López**
Universidad de Jaén, Campus Las Lagunillas
Edificio A3. E-23071
{jmperea,magc,mgarcia,laurena}@ujaen.es

Resumen: En este artículo se presenta un estudio comparativo de las distintas estrategias y técnicas de procesamiento del lenguaje natural más utilizadas en la actualidad para abordar la tarea de la recuperación de información geográfica (*Geographical Information Retrieval*, GIR). Este trabajo se ha basado fundamentalmente en el análisis de los mejores sistemas presentados a la tarea de búsqueda del GeoCLEF, un marco de evaluación para recuperación de información geográfica que pertenece al foro internacional *Cross Language Evaluation Forum* (CLEF). Las conclusiones obtenidas reflejan que es imprescindible hacer uso de recursos externos de información geográfica, tales como *gazetteers* y tesauros o reconocedores de entidades. Así mismo es necesario realizar una indexación por separado de la información geográfica y de la no geográfica antes del proceso de recuperación.
Palabras clave: Recuperación de Información Geográfica, GeoCLEF, Procesamiento del Lenguaje Natural, Recuperación de Información

Abstract: This paper presents a comparative study of several strategies and techniques of natural language processing most used at present to solve the geographical retrieval information (GIR) task. This work has been based on the analysis of the best systems submitted to the search task of GeoCLEF, an evaluation framework for the geographical information retrieval task which belongs to the international forum Cross Language Evaluation Forum (CLEF). The main conclusions show that it is imperative to make use of external geographic information resources such as gazetteers and thesaurus, named entity recognizers and it is necessary to make an index for geographic information only and another index for non-geographic information before the retrieval process.
Keywords: Geographical Information Retrieval, GeoCLEF, Natural Language Processing, Information Retrieval

1. Introducción

La recuperación de información geográfica (GIR a partir de ahora, del inglés *Geographical Information Retrieval*) pertenece a una rama especializada de la recuperación de información (IR, del inglés *Information Retrieval*) tradicional. Incluye todas las áreas de investigación que tradicionalmente forman el núcleo de la IR, pero además con un énfasis en la información geográfica y espacial. La recuperación de información geográfica se preocupa de la recuperación de información que involucra algún tipo de percepción espacial. Muchos documentos contienen algún tipo de referencia espacial relevante para la búsqueda (Mandl et al., 2007).

Existen congresos y foros de evaluación como el Text REtrieval Conference[1] (TREC) y el CLEF[2] que no evalúan expresamente la relevancia en la tarea de la recuperación de información geográfica. El objetivo del Geo-

* Este trabajo ha sido financiado por el Ministerio de Ciencia y Tecnología a través del proyecto TIMOM (TIN2006-15265-C06-03) y el proyecto RFC/PP2006/Id514 financiado por la Universidad de Jaén.

[1] http://trec.nist.gov
[2] http://www.clef-campaign.org

 © Sociedad Española para el Procesamiento del Lenguaje Natural

CLEF[3] es proporcionar el marco de trabajo necesario en el que evaluar estos sistemas GIR en búsquedas de información, teniendo en cuenta aspectos geo-referenciales y multilingües. Es una tarea perteneciente al CLEF que se viene celebrando desde 2005.

La principal contribución de este artículo es ofrecer una visión general de las estrategias y técnicas de procesamiento del lenguaje natural (PLN) más utilizadas en los sistemas presentados a la tarea GeoCLEF durante los últimos tres años, para resolver la recuperación de información basada en contenido geográfico. El artículo se organiza de la siguiente manera: en primer lugar, se describe brevemente la tarea de la recuperación de información geográfica. A continuación, se presentan los recursos utilizados en GeoCLEF. Las principales estrategias usadas en un sistema de recuperación de información geográfica se describen en la siguiente sección. En la sección cinco se muestra un análisis de los resultados obtenidos en el marco del GeoCLEF. Finalmente, se comentan las conclusiones.

2. *La tarea de la recuperación de información geográfica*

Se puede definir la tarea de la recuperación de información geográfica como la recuperación de documentos relevantes en respuesta a una consulta con el formato <tema, localización>, donde la relación espacial puede implicar implícitamente contenido, o explícitamente ser seleccionado de un conjunto de posibles opciones topológicas, direccionales o de proximidad (Bucher et al., 2005).

La tarea más importante definida en GeoCLEF es la de búsqueda de información geográfica (*search task*). Pero GeoCLEF no sólo evalúa sistemas de búsqueda de información geográfica, sino que también está proponiendo nuevas subtareas que se enmarcan dentro de esta rama, como la de análisis de consultas (*query parsing*), cuyo objetivo es identificar aspectos geográficos en una consulta, o las subtareas piloto que han propuesto para este año 2008 relacionadas con Wikipedia[4] y la búsqueda geográfica de imágenes. Para la tarea principal de búsqueda, GeoCLEF organiza a su vez dos subtareas: la monolingüe,

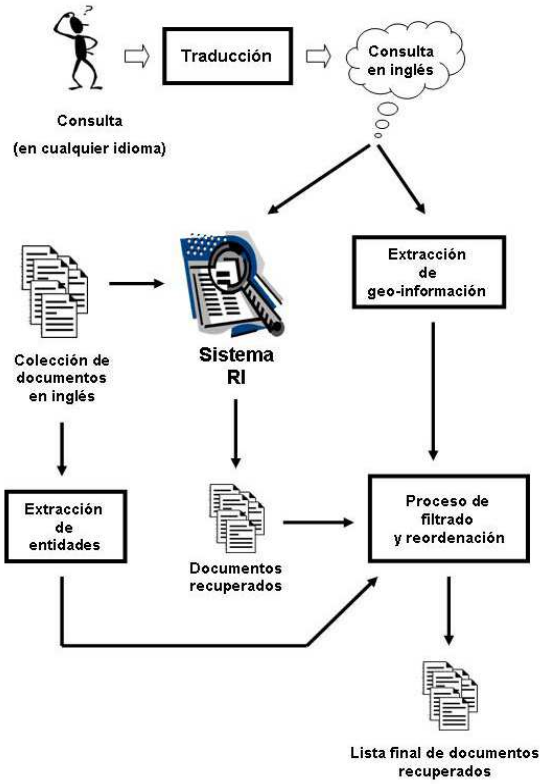

Figura 1: Arquitectura básica del sistema GIR GeoUJA

en la que hay que utilizar el mismo idioma tanto para las consultas como para las colecciones (inglés, alemán o portugués en 2007), y la bilingüe, que implica traducción, ya que el idioma de la consulta tiene que ser distinto al de la colección utilizada.

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

En la Figura 1 se puede observar la arquitectura básica empleada en el sistema GIR *GeoUJA* (Perea Ortega et al., 2007). Este sistema ha sido desarrollado por nuestro grupo de investigación SINAI[5] para resolver la tarea

[3] http://ir.shef.ac.uk/geoclef
[4] http://www.wikipedia.org
[5] http://sinai.ujaen.es

de la recuperación de información geográfica, presentando distintas versiones del mismo en las competiciones de GeoCLEF 2006 (García Vega et al., 2007) y 2007.

3. Recursos

Las colecciones de documentos utilizadas en GeoCLEF constan de relatos periodísticos ocurridos en los años 1994 y 1995. La colección de inglés contiene historias, noticias y eventos de cobertura nacional e internacional que representan una amplia variedad de regiones geográficas y localizaciones. Esta colección consta de un total de 169.477 documentos y fue compuesta con noticias del periódico inglés *The Glasgow Herald* (1995) y del periódico americano *Los Angeles Times* (1994). Además de la colección en inglés, GeoCLEF 2007 proporcionó colecciones en idioma alemán y portugués. En GeoCLEF 2006 se llegó a facilitar incluso una colección de documentos en español. Todas estas colecciones tienen una estructura común: información específica de periódico como fecha, página, tema, título, autor y el texto de la noticia. Las colecciones no han sido etiquetadas geográficamente y no contienen información semántica específica sobre localizaciones (Mandl et al., 2007).

Un total de 25 consultas fueron generadas para GeoCLEF 2007. Estas consultas han intentado reflejar un punto de vista de usuario razonable, bien preguntando por lugares turísticos (por ejemplo la catedral de *St. Paul*), definiendo zonas específicas ("*al norte de Italia*"), o bien desde un punto de vista periodístico ("*violación de derechos humanos en Myanmar*" o "*muertes en el Himalaya*"). También se han tratado de reflejar distintas dificultades relacionadas con tareas que aborda el procesamiento del lenguaje natural:

- Ambigüedad geográfica. Por ejemplo, existe una catedral de *St. Paul* en Londres y otra en Sao Paulo.
- Regiones geográficas mal definidas ("*cerca del este*").
- Relaciones geográficas complejas como "*cerca de ciudades rusas*" o "*a lo largo de la costa mediterránea*".
- Aspectos multilingües. "*Greater Lisbon*" en inglés es lo mismo que "*grande Lisboa*" en portugués o que "*großraum Lissabon*" en alemán.
- Granularidad en las referencias a países. Por ejemplo, "*al norte de Italia*".

El formato utilizado para las consultas en los años 2006 y 2007 difiere ligeramente del empleado en 2005, ya que no proporciona las entidades geográficas ya etiquetadas.

Como se puede observar en la Figura 2, una consulta consta de tres etiquetas: título (<*title*>), descripción (<*desc*>) y narrativa (<*narr*>). Normalmente para los experimentos se suele utilizar el texto de las etiquetas título y descripción, aunque para algunas consultas es interesante usar el texto de la etiqueta narrativa, ya que contiene descripciones geográficas detalladas que ayudan al motor de búsqueda a definir con más exactitud su criterio de relevancia e incluso, a veces, contiene listados de localizaciones o regiones relevantes para la búsqueda.

```
<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 2: Formato de una consulta del GeoCLEF 2007

4. Principales técnicas de PLN aplicadas en un sistema GIR

En el estudio de las principales técnicas PLN aplicadas en una arquitectura GIR nos hemos basado en los sistemas presentados en GeoCLEF 2005, 2006 y 2007 para la tarea monolingüe en inglés.

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

Según el estudio realizado, el stemmer más utilizado es el ***Porter Stemmer***[6]. También

[6] http://tartarus.org/martin/PorterStemmer

se usa en varios sistemas, pero con menos frecuencia que el anterior, el ***Snowball Tartarus***[7]. Con respecto a la lista de ***stopwords*** para el inglés, la más utilizada ha sido la creada por Salton y Buckley[8], que consta de 571 palabras. En relación a los reconocedores de entidades más empleados, hay sistemas que han optado por implementar sus propios reconocedores haciendo uso de distintas bases de conocimiento geográficas y tesauros (Ferrés y Rodríguez, 2007), (Larson, 2007), pero la mayoría han empleado ***Lingpipe***[9] como herramienta NER. En nuestro sistema GIR presentado a las dos últimas ediciones del GeoCLEF hemos hecho uso del módulo NER que incorpora la herramienta GATE (*General Architecture for Text Engineering*)[10], obteniendo buenos resultados.

Según el análisis de los distintos sistemas, es poco habitual utilizar herramientas de etiquetado POS (*Part Of Speech*), aunque algunos sistemas como (Ferrés y Rodríguez, 2007) hacen uso de un etiquetador POS estadístico llamado *TnT*.

Por último, otra herramienta importante en el ámbito del PLN son los traductores o sistemas de traducción automática (*Machine Translation*, MT). Para la tarea GIR es necesario utilizarlos cuando la consulta planteada y la colección a indexar están en idiomas distintos (tarea multilingüe). En (Larson, 2007) se hace uso del traductor LEC Power Translator. En nuestro sistema GIR *GeoUJA* utilizamos un sistema propio de traducción automática llamado SINTRAM (SINai TRAnslation Module) (García Cumbreras et al., 2007).

5. Aproximaciones más utilizadas para resolver la tarea GIR

En general, la arquitectura de cualquier sistema GIR parte de un modelo básico de recuperación de información. Por tanto, un elemento esencial en todos los sistemas presentados es la herramienta utilizada como motor de búsqueda. Entre los más usados están **Lucene**[11], **Terrier**[12] y algo menos **Lemur**[13]. Algunos participantes han optado por implementar su propio motor de búsqueda, como en (Toral et al., 2006), con el sistema *IR-n*, basado en pasajes, obteniendo buenos resultados en la competición GeoCLEF 2006.

Según el estudio, los esquemas de pesado más utilizados en los sistemas IR han sido: **TF·IDF**, **Okapi** (Robertson y Walker, 1999), **DFR** (*Divergence From Randomness*) (Ounis et al., 2006), **BRF** (*Blind Relevance Feedback*) (Chen, 2003), **PRF** (*Pseudo Relevant Feedback*) (Buckley et al., 1995) y **LR** (*Logistic Regression*) o modelo de Regresión Logística (Cooper, Gey, y Dabney, 1992). Existen otros esquemas menos usuales como el de frecuencia inversa de documento con normalización 2 de Laplace o ***InL2***, utilizado en (Guillén, 2007).

5.1. GeoCLEF 2005

En la primera edición del GeoCLEF, a diferencia de las dos posteriores, los organizadores añadieron en las consultas información sobre el concepto principal, las localizaciones y las relaciones espaciales de las mismas. Toda esta información fue extraída de forma manual y colocada en etiquetas justo después de las principales de cada *topic*.

Por este motivo, hubo algunas aproximaciones basadas únicamente en recuperación de información clásica, sin ningún tratamiento geográfico. De hecho, de los cuatro sistemas con mayor puntuación en esta edición, tres de ellos se basaron únicamente en un sistema de IR sin tratamiento de la información geográfica. La arquitectura que obtuvo mejores resultados en la tarea monolingüe de inglés fue la presentada por la Universidad de Berkeley (Gey y Petras, 2005), que utilizó un sistema clásico de recuperación de información con un algoritmo de ranking de documentos basado en regresión logística.

La mayoría de sistemas apostaron por utilizar reconocedores de entidades especializados en el dominio geográfico como una aproximación inicial para resolver esta tarea (Cardoso et al., 2005). Otras arquitecturas también emplearon recursos externos de conocimiento geográfico tales como ontologías y *gazetteers*, así como estadísticas sociales y características físicas de los mismos. En concreto, hicieron uso de gazetteers como **GNIS**[14] (Geographic Names Information System) y **GNS**[15] (Geonet Names Ser-

[7] http://snowball.tartarus.org
[8] ftp://ftp.cs.cornell.edu/pub/smart/english.stop
[9] http://www.alias-i.com/lingpipe
[10] http://gate.ac.uk
[11] http://lucene.apache.org
[12] http://ir.dcs.gla.ac.uk/terrier
[13] http://www.lemurproject.org
[14] http://www.usgs.gov
[15] http://www.nga.mil

ver). El grupo XLDB de la Universidad de Lisboa construyó su propia ontología geográfica basándose en recursos externos como Wikipedia y **World Gazetteer**[16] (Cardoso et al., 2005).

Por otro lado, hubo varios sistemas que utilizaron expansión de consulta (Buscaldi, Rosso, y Sanchis Arnal, 2005). La arquitectura presentada por la Universidad Politécnica de Valencia hizo uso de la ontología no geográfica WordNet[17] para realizar dicha expansión, basándose en las relaciones de sinonimia y meronimia.

5.2. GeoCLEF 2006

En GeoCLEF 2006 la variación de arquitecturas presentadas en los distintos sistemas aumentó considerablemente con respecto a la primera edición. Estas aproximaciones variaban desde enfoques básicos de IR sin indexación geográfica a profundos procesamientos del lenguaje natural para extraer lugares y términos topológicos tanto de las colecciones como de las consultas. Algunas de las técnicas específicas usadas fueron:

- Técnicas ad-hoc (BRF, descomposición de palabras, expansión manual de consultas).
- Construcción propia de recursos de conocimiento geográfico a partir de recursos externos (*gazetteers* como GNIS o World Gazetteer).
- Expansión de consultas basada en *gazetteer* y WordNet.
- Módulos de pregunta-respuesta utilizando recuperación de pasajes.
- Extracción de entidades geográficas.
- Resolución de la ambigüedad geográfica.

El sistema presentado por el grupo XLDB de la Universidad de Lisboa (Martins et al., 2006) fue el que obtuvo mejores resultados en la tarea monolingüe en inglés. Volvieron a hacer uso de la ontología geográfica que crearon en la edición anterior y la utilizaron para expandir las consultas. Esta ontología se organiza en conceptos que ellos hacen corresponder con ámbitos geográficos (*geographic scopes*). De este modo, también utilizaron expansión de consultas basadas en ámbitos geográficos. Otra característica interesante de su sistema es que hicieron uso de desambiguación de referencias geográficas (topónimos) y de similitud geográfica entre ámbitos.

Nuestro grupo de investigación SINAI, en su primera participación en GeoCLEF (García Vega et al., 2007), optó por el enfoque de expandir las consultas utilizando información geográfica procedente de un NER, de un gazetteer como **Geonames**[18] y de un tesauro generado a partir de las propias colecciones del GeoCLEF. Esta aproximación no ofreció mejores resultados que el caso base (sin expansión de consultas) por lo que concluimos que la expansión no se estaba haciendo correctamente. Esto mismo le ocurrió a la Universidad de Alicante, que quedó en segunda posición en la tarea monolingüe en inglés. El enfoque básico que utilizó este grupo fue el que siguieron la mayoría de sistemas presentados en esta segunda edición del GeoCLEF (Toral et al., 2006).

5.3. GeoCLEF 2007

GeoCLEF 2007 se presentaba con la novedad de una nueva tarea: clasificación de consultas. Su objetivo era identificar componentes geográficos en las mismas. La tarea principal mantuvo las subtareas monolingüe y bilingüe. Los organizadores continuaron con su esfuerzo de proponer un conjunto de consultas difíciles desde el punto de vista geográfico (ver apartado 3).

El mejor sistema en la tarea de búsqueda monolingüe en inglés fue el presentado por la Universidad Politécnica de Cataluña (Ferrés y Rodríguez, 2007). En este enfoque, a partir del texto de las colecciones, se construyen dos índices:

- **Índice geográfico**. Contiene toda la información geográfica extraída del texto de las colecciones (entidades, variaciones de nombres de entidades para resolver posibles ambigüedades, coordenadas geográficas, etc.).
- **Índice textual**. Almacena los lemas de las palabras con contenido semántico de la colección, sin incluir ninguna información geográfica.

Para extraer la información geográfica tanto de las colecciones como de las consultas, hacen uso de una **base de conocimien-**

[16] http://world-gazetteer.com
[17] http://wordnet.princeton.edu
[18] http://www.geonames.org

to geográfico generada por ellos mismos y que consta de tres componentes:

- Un tesauro geográfico. Este componente fue construido a su vez uniendo cuatro *gazetteers*: GNS, GNIS, *GeoWorldMap*[19] y World Gazetteer. Como cada gazetteer tiene distintas clases y conceptos, ellos mapearon estas clases al conjunto de características proporcionado por el tesauro ADL Feature Type Thesaurus[20] (ADLFTT).
- Un tesauro de tipos de características. Utilizaron el tesauro ADL Feature Type Thesaurus.
- Una base de datos que contiene conjuntos de regiones no coincidentes (representadas por polígonos) para cada país (Pouliquen et al., 2004). Esta base de datos resuelve tareas como la obtención de los límites de cualquier país, la detección de si unas coordenadas dadas pertenecen a una determinada área, etc.

Antes del proceso de recuperación, una fase importante en este sistema es el análisis de la consulta. Este procesamiento se divide en un **análisis lingüístico** de los *topics* (etiquetado POS, extracción de lemas y de entidades) y en un **análisis geográfico**, aplicado sobre las localizaciones y organizaciones detectadas durante el análisis lingüístico, y que hace uso de la base de conocimiento geográfica explicada anteriormente.

Con todos estos ingredientes lanzan la recuperación de documentos teniendo como consulta los lemas (sin información geográfica) del topic en cuestión. Para ello, utilizan Terrier como motor de búsqueda con varios esquemas de pesado (TF·IDF, Okapi y DFR). Por otro lado, obtienen otra lista de documentos recuperados utilizando la información geográfica extraída del topic y el índice geográfico creado con anterioridad. Como motor de búsqueda en este índice hacen uso de un sistema IR basado en pregunta-respuesta (*Question-Answering based IR system*).

La última fase de la arquitectura consta de un proceso de filtrado con los documentos recuperados por el sistema IR y los recuperados usando la base de conocimiento geográfico y el índice geográfico. En el ranking final de documentos se colocan primero aquellos que aparezcan en las dos listas. Se puede ver un esquema del enfoque seguido por la Universidad Politécnica de Cataluña en la Figura 3.

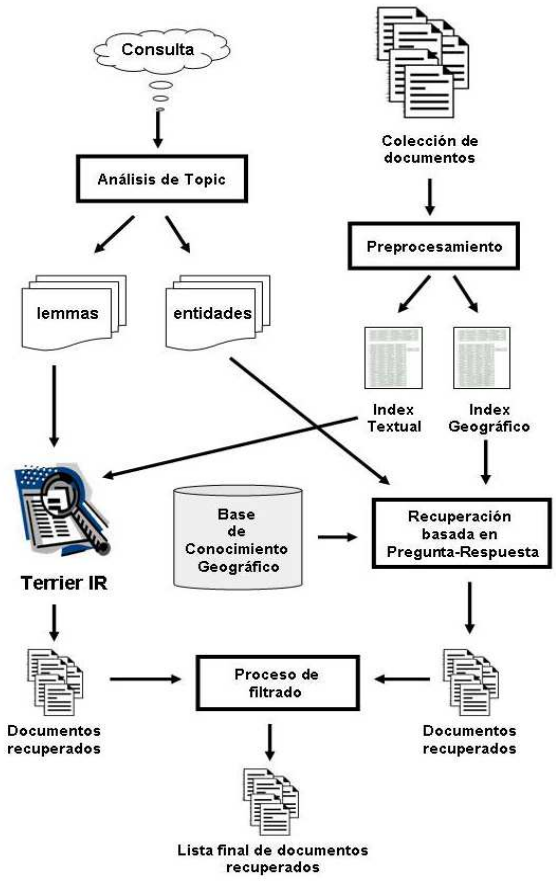

Figura 3: Arquitectura básica del sistema TALP presentado por la Universidad Politécnica de Cataluña en GeoCLEF 2007

El resto de sistemas presentados optaron básicamente por la misma filosofía de usar recursos geográficos externos, *gazetteers*, tesauros, ontologías como WordNet e incluso Wikipedia. Mencionar la propuesta de la Universidad Politécnica de Valencia (Buscaldi y Rosso, 2007) que utilizó expansión de consultas con WordNet haciendo uso de tres índices: uno para términos geográficos (topónimos); otro para términos no geográficos y el último para términos extraídos de WordNet holónimos y sinónimos de los topónimos encontrados en el primer índice.

6. Análisis de resultados

En esta sección vamos a analizar los resultados obtenidos por los distintos participantes de las tres últimas ediciones del GeoCLEF para la tarea monolingüe en inglés (ver

[19] http://www.geobytes.com
[20] http://www.alexandria.ucsb.edu/gazetteer

Año	Universidad	MAP
2005	Berkeley2	**0.3936**
2005	San Marcos	0.3613
2005	Alicante	0.3495
2006	Lisboa	**0.3034**
2006	Alicante	0.2723
2006	San Marcos	0.2637
2007	Politécnica Cataluña	**0.2850**
2007	Berkeley1	0.2642
2007	Politécnica Valencia	0.2636

Tabla 1: Principales resultados del GeoCLEF en la tarea monolingüe inglés

Tabla 1).

En general, se observa una decadencia de resultados en términos de precisión media (*Mean Average Precision*, MAP) desde 2005 a 2007. Esto es debido fundamentalmente a la mayor innovación y diversidad introducida a la hora de generar las consultas tanto del 2006 como del 2007. Por ejemplo, para los topics del GeoCLEF 2007 se introdujeron dificultades añadidas como relaciones geográficas complejas ("*la costa mediterránea*"), regiones políticas ("*Bosphorus*") o lugares geográficos delicados como lagos, aeropuertos, circuitos de fórmula uno o catedrales. Todo esto ha hecho que la dificultad en resolver la tarea aumente y la precisión obtenida por los sistemas empeore.

7. Conclusiones

En este trabajo se ha presentado un estudio sobre las distintas estrategias empleadas para resolver la tarea de la recuperación de información geográfica (GIR), así como las técnicas de PLN más utilizadas. Dicho estudio se ha centrado en los sistemas presentados en GeoCLEF, un marco de evaluación GIR que organiza el CLEF desde el año 2005. Las conclusiones que se derivan de este estudio se resumen a continuación:

- Es imprescindible hacer uso de recursos externos de información geográfica, tales como *gazetteers* y tesauros. Algunos de los más utilizados son: GNIS, GNS, Geonames, World Gazetteer o GeoWorldMap.
- Es recomendable la creación de al menos dos índices para el proceso de recuperación de información: uno que contenga la información no geográfica (índice textual) y otro sólo con la información geográfica (entidades, georeferencias, relaciones espaciales, etc.).
- Técnicas PLN básicas aplicadas tanto a las colecciones como a las consultas: detector y reconocedor de entidades (NER), lematizador, lista de palabras vacías y etiquetador POS.
- Sería interesante contar también con un desambigüador de topónimos para resolver ambigüedades geográficas.
- En cuanto a la expansión de consultas no queda claro si es recomendable utilizarla. Hay sistemas que han empeorado sus resultados usando esta técnica como (García Vega et al., 2007) o (Toral et al., 2006) y otros que los han mejorado (Buscaldi y Rosso, 2007) o (Ferrés y Rodríguez, 2007).
- El uso de otros recursos como WordNet o Wikipedia también pueden ser interesantes.

Bibliografía

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

Buckley, C., G. Salton, J. Allan, y A. Singhal. 1995. Automatic query expansion using smart: Trec 3. *Proceedings of TREC3. NIST, Gaithesburg, MD*, páginas 69–80.

Buscaldi, D. y P. Rosso. 2007. The UPV at GeoCLEF 2007. En *Working Notes of the Cross Language Evaluation Forum (CLEF 2007)*.

Buscaldi, D., P. Rosso, y E. Sanchis Arnal. 2005. A WordNet-based Query Expansion method for Geographical Information Retrieval. En *Working Notes of the Cross Language Evaluation Forum (CLEF 2005)*.

Cardoso, N., B. Martins, M. Silveira Chaves, L. Andrade, y M.J. Silva. 2005. The XLDB Group at GeoCLEF 2005. En *Working Notes of the Cross Language Evaluation Forum (CLEF 2005)*.

Chen, Aitao. 2003. *Cross-Language Retrieval Experiments at CLEF 2002*, volumen 2785 of LNCS Series. Springer-Verlag.

Cooper, W.S., F.C. Gey, y D.P. Dabney. 1992. Probabilistic retrieval based on staged logistic regression. En *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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

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

García Vega, M., M.A. García Cumbreras, L.A. Ureña López, y J.M. Perea Ortega. 2007. *GEOUJA System. The first participation of the University of Jaén at GEOCLEF 2006*, volumen 4730 of LNCS Series. Springer-Verlag.

Gey, F. y V. Petras. 2005. Berkeley2 at GeoCLEF: Cross-Language Geographic Information Retrieval of German and English Documents. En *Working Notes of the Cross Language Evaluation Forum (CLEF 2005)*.

Guillén, R. 2007. GeoCLEF2007 Experiments in Query Parsing and Cross-language GIR. En *Working Notes of the Cross Language Evaluation Forum (CLEF 2007)*.

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

Mandl, T., F. Gey, Di Nunzio, G., N. Ferro, R. Larson, M. Sanderson, D. Santos, C. Womser-Hacker, y Xing Xie. 2007. Geoclef 2007: the clef 2007 cross-language geographic information retrieval track overview. En *Proceedings of the Cross Language Evaluation Forum (CLEF 2007)*.

Martins, B., N. Cardoso, M. Silveira Chaves, L. Andrade, y M.J. Silva. 2006. The University of Lisbon at GeoCLEF 2006. En *Working Notes of the Cross Language Evaluation Forum (CLEF 2006)*.

Ounis, I., G. Amati, V. Plachouras, B. He, C. Macdonald, y C. Lioma. 2006. Terrier: A High Performance and Scalable Information Retrieval Platform. En *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006). Seattle, Washington, USA*.

Perea Ortega, J.M., M.A. García Cumbreras, M. García Vega, y A. Montejo Ráez. 2007. GEOUJA System. University of Jaén at GEOCLEF 2007. En *Working Notes of the Cross Language Evaluation Forum (CLEF 2007)*, página 52.

Pouliquen, B., R. Steinberger, C. Ignat, y T. De Groeve. 2004. Geographical information recognition and visualization in texts written in various languages. En *Proceedings of the 2004 ACM symposium on Applied computing*, páginas 1051–1058.

Robertson, S.E. y S. Walker. 1999. Okapi-Keenbow at TREC-8. En *Proceedings of the 8th Text Retrieval Conference TREC-8, NIST Special Publication 500-246*, páginas 151–162.

Toral, A., O. Ferrández, Noguera, E., Z. Kozareva, A. Montoyo, y R. Muñoz. 2006. Geographic IR Helped by Structured Geospatial Knowledge Resources. En *Working Notes of the Cross Language Evaluation Forum (CLEF 2006)*.