Procesamiento del Lenguaje Natural, Revista nº 50 marzo de 2013, pp 69-76 recibido 30-11-12 revisado 11-02-13 aceptado 19-02-13

Una aproximación basada en corpus para la detección del foco geográfico en el texto

A corpus-based approach to geographical focus detection in text

Fernando S. Peregrino, David Tomás, Fernando Llopis
Universidad de Alicante
Carretera San Vicente del Raspeig s/n - 03690 Alicante (Spain)
{fsperegrino, dtomas, llopis}@dlsi.ua.es

Resumen: El foco geográfico de un documento identifica el lugar o lugares en los que se centra el contenido del texto. En este trabajo se presenta una aproximación basada en corpus para la detección del foco geográfico en el texto. Frente a otras aproximaciones que se centran en el uso de información puramente geográfica para la detección del foco, nuestra propuesta emplea toda la información textual existente en los documentos del corpus de trabajo, partiendo de la hipótesis de que la aparición de determinados personajes, eventos, fechas e incluso términos comunes, pueden resultar fundamentales para esta tarea. Para validar nuestra hipótesis, se ha realizado un estudio sobre un corpus de noticias geolocalizadas que tuvieron lugar entre los años 2008 y 2011. Esta distribución temporal nos ha permitido, además, analizar la evolución del rendimiento del clasificador y de los términos más representativos de diferentes localidades a lo largo del tiempo.
Palabras clave: Foco geográfico, recuperación de información geográfica, clasificación de textos, aprendizaje automático

Abstract: The geographical focus of a document identifies the relevant locations mentioned in text. This paper presents a corpus-based approach to detecting the geographical focus in documents. Despite other approaches focused on using solely geographical information, our proposal employs all the textual information included in the corpus under the assumption that the presence of particular names of persons, events, and even common terms can definitely help to solve this task. In order to validate our hypothesis, a study was carried out on a corpus of georeferenced news that took place between 2008 and 2011. Moreover, this temporal distribution allowed to carry out a study on the evolution of the performance of the classifier and the most representative terms for different locations over time.
Keywords: Geographical focus, geographical information retrieval, text classification, machine learning

1 Introducción

La identificación del foco geográfico de un documento consiste en determinar la principal o principales localizaciones a las que se hace referencia en el texto de entre todas las que se nombran en él (Amitay et al., 2004). Por ejemplo, en un documento que trata temas relativos a la Comunidad Valenciana, pueden aparecer frases como "*La compañía proveniente de China ha establecido en Valencia su segunda mayor fábrica, después de la de San Francisco, junto con otras empresas locales.*". Dicho documento, pese a citar localizaciones como China y San Francisco, tiene como único foco geográfico la Comunidad Valenciana, siendo la relevancia del resto de entidades meramente testimonial.

Obtener el foco geográfico de un documento es fundamental a la hora de desarrollar sistemas de *recuperación de información geográfica* (*Geographic Information Retrieval*, GIR). Estos sistemas son una especialización de los sistemas tradicionales de *recuperación de información* enfocados a la obtención de documentos relevantes para una determinada localización. Identificar el foco geográfico puede permitir a estos sistemas determinar con mayor precisión la relevancia

ISSN 1135-5948 © 2013 Sociedad Española para el Procesamiento del Lenguaje Natural

que tiene un documento para un determinado lugar, descartando aquellos documentos donde la mención de una localización es puntual e irrelevante. Más aún, en ocasiones el foco geográfico ni siquiera se menciona explícitamente en el texto. En estos casos, la presencia de otros elementos geográficos que se nombran en él pueden ayudar a inferirlo (Amitay et al., 2004).

El presente trabajo se centra en el problema de la detección del foco geográfico en el texto. Esta tarea se ha abordado desde el punto de vista de la clasificación textual. Empleando un corpus de noticias previamente geolocalizadas y pertenecientes a 61 localidades del estado Español, hemos entrenado un sistema de clasificación automático capaz de determinar, para una nueva noticia, el principal foco de atención geográfico tratado en ella, asignando la localidad más probable de entre las 61 posibles. A diferencia de aproximaciones anteriores centradas únicamente en la información geográfica, nuestra propuesta emplea todo el contenido textual de los documentos, utilizando de esta manera un conocimiento general del mundo para la clasificación. En nuestro caso, partimos de la hipótesis de que las personas, eventos, fechas y términos comunes que caracterizan una localidad pueden resultar de gran utilidad a la hora de determinar su preponderancia en el texto.

Adicionalmente, dado que el corpus de noticias empleado en este estudio pertenece a diferentes años naturales (desde 2009 hasta 2011), hemos realizado un estudio de la evolución del rendimiento del clasificador a lo largo del tiempo, así como de la evolución de los términos más representativos de alguna de las localidades existente en el corpus.

El resto de este artículo está organizado del siguiente modo: en la Sección 2, se comentan los trabajos más destacados dentro del campo de la identificación del foco geográfico en el texto; la Sección 3 describe el corpus utilizado en nuestros experimentos; en la Sección 4 se describen los experimentos y se muestran los resultados que se han llevado a cabo; la Sección 5 resume las conclusiones de este estudio y las propuestas de trabajo futuro.

2 Trabajo relacionado

Los documentos de texto se pueden asociar frecuentemente con un determinado contexto geográfico. Para la detección de las localizaciones en el texto es común el uso de técnicas de *reconocimiento de entidades nombradas* (*Named Entity Recognition*) capaces de detectar entes de tipo geográfico. Esta detección va siempre asociada a un proceso posterior de desambiguación, a fin de concretar a cuál de las múltiples localizaciones a las que se puede asociar un mismo nombre se está refiriendo el texto (Buscaldi y Rosso, 2008). Sirvan como ejemplo los 18 Jerusalem y 63 Springfields que hay en Estados Unidos.

Este proceso de detección, sin embargo, no proporciona información sobre la verdadera relevancia que tienen las entidades geográficas nombradas en el texto. Si bien el tema de la detección y desambiguación de topónimos es un tema ampliamente tratado en la literatura científica (Leidner, 2007), son escasos los trabajos centrados en determinar el grado real de relevancia de las entidades geográficas que aparecen en el texto.

En (Martins y Silva, 2005) los autores presentan una variante del algoritmo *PageRank* para la detección del ámbito geográfico de cada documento. Para ello usan las referencias geográficas extraídas del texto y una técnica basada en ontologías, combinando ambas e infiriendo el ámbito global de cada documento a partir de la aplicación de *PageRank* al grafo generado a partir de la ontología.

En otro trabajo de los mismos autores (Anastácio, Martins, y Calado, 2009a), un conjunto de documentos obtenidos de la Web son categorizados en función de las localizaciones expresadas en ellos, utilizando *máquinas de vectores de soporte* (*Support Vectors Machines*, SVM) con diferentes vectores de características, como por ejemplo los n-gramas extraídos de las URLs de las páginas web.

El estudio presentado en (Ye et al., 2011) se centra en identificar la localización de la que se habla en un conjunto de blogs de viaje. Para ello, los autores implementan un extractor de localizaciones (Qin et al., 2010) con el fin de obtener los lugares mencionados en estos blogs. Debido a la inherente ambigüedad de los topónimos, exploran rasgos textuales en su contexto cercano (palabras alrededor del topónimo identificado) y rasgos geográficos (relaciones geográficas entre las localizaciones del texto) para llevar a cabo la clasificación por relevancia de cada una de las localizaciones identificadas.

En (Amitay et al., 2004), los autores geolocalizan contenido de la Web mediante el uso de diccionarios geográficos (*gazetteers*). Estos diccionarios son empleados para la desambiguación de topónimos, a los que se les asigna un factor de confianza. Este factor de confianza, junto con la frecuencia de aparición del topónimo en el texto y el resto de entidades geográficas que tienen lugar en él, determinan un valor final de relevancia que permite identificar el foco geográfico del texto. La novedad de esta aproximación es la capacidad de detectar el foco de un documento incluso cuando éste no aparece explícitamente en él, infiriéndolo a partir del resto de localizaciones presentes.

Finalmente, en (Anastácio, Martins, y Calado, 2009b) se pude ver un estudio comparativo sobre cuatro sistemas utilizados para la asignación del foco geográfico: *Yahoo! PlaceMaker*,[1] *Web-a-Where* (Amitay et al., 2004), *GIPSY* (Woodruff y Plaunt, 1994) y *GREASE* (Martins y Silva, 2005). La aproximación vencedora fue la llevada a cabo por el sistema *Web-a-Where*.

Por lo que respecta a aplicaciones finales de estas tecnologías, en (Clough et al., 2011) los autores muestran como enlazar los archivos oficiales del gobierno británico (*UK National Archives*) con el foco geográfico de los documentos para mejorar el acceso a los mismos.

A diferencia de la mayoría de sistemas aquí mencionados, nuestra propuesta no se centra únicamente en la información geográfica para determinar el foco, sino que emplea toda la información textual presente en un documento para su identificación. Esto nos va a permitir identificar el foco geográfico incluso en las situaciones en las que éste no aparece de forma explícita en el texto, ya que el resto de información presente en él nos puede llevar a inferir sobre qué localidad se está hablando.

3 Corpus de trabajo

Para poder llevar a cabo los experimentos propuestos en este trabajo, recopilamos un conjunto de noticias locales en español pertenecientes al periódico *20 Minutos*.[2] El corpus resultante consistió en más de 500.000 noticias (ver Tabla 1), comprendidas entre los años 2008 y 2011, pertenecientes a 61 localidades de España.[3]

Año	Noticias	Vocabulario
2.008	55.019	182.716
2.009	29.394	145.191
2.010	224.729	370.007
2.011	224.179	377.153
Total	553.321	606.229

Tabla 1: Estadísticas del corpus recopilado. Se muestra el número de noticias y el tamaño del vocabulario para cada uno de los años.

El hecho de que cada noticia tenga asociada una localidad en el corpus, nos permite considerar a dicha localidad como el foco geográfico de la noticia, ya que su contenido ha sido considerado como relevante para esa localización concreta. Este corpus de gran tamaño nos va a permitir aplicar técnicas de aprendizaje automático para construir un clasificador capaz de asignar un foco geográfico, para una nueva noticia dada, de entre las 61 localidades mencionadas anteriormente.

4 Experimentos y resultados

Para evaluar nuestra hipótesis de partida, hemos realizado diversos experimentos sobre el corpus descrito en la sección anterior. En primer lugar, hemos evaluado diferentes algoritmos de clasificación para determinar aquél que nos pudiera proporcionar un mejor rendimiento en la tarea.

El segundo experimento ha consistido en un estudio sobre la selección de características para reducir el número de dimensiones de las instancias de aprendizaje. El objetivo es incrementar el rendimiento del clasificador eliminando ruido (características innecesarias), reduciendo además el tiempo de computación necesario para llevar a cabo el proceso de entrenamiento y evaluación.

[1] http://developer.yahoo.com/geo/placemaker/

[2] http://www.20minutos.es/

[3] A Coruña, Albacete, Algeciras, Alicante, Almeria, Ávila, Badajoz, Barcelona, Bilbao, Burgos, Cáceres, Cádiz, Cartagena, Castellón de la Plana, Ceuta, Ciudad Real, Córdoba, Cuenca, Elche, Gijón, Girona, Granada, Guadalajara, Huelva, Huesca, Jaén, Jerez de la Frontera, Las Palmas de Gran Canaria, León, Lleida, Logroño, Lugo, Madrid, Málaga, Marbella, Melilla, Ourense, Oviedo, Palencia, Palma de Mallorca, Pamplona, Pontevedra, Salamanca, San Sebastián, Santa Cruz de Tenerife, Santander, Santiago de Compostela, Segovia, Sevilla, Soria, Teruel, Toledo, Valencia, Valladolid, Vigo, Vitoria, Zamora y Zaragoza.

El tercer experimento se ha enfocado a evaluar la evolución del rendimiento del clasificador al ser entrenado con muestras de un año (2008) y evaluado con muestras de años posteriores (2009 a 2011). Esto nos va a permitir ver cómo la evolución del vocabulario, un concepto muy vivo en el caso de las noticias de prensa, afecta al rendimiento del sistema con el paso del tiempo.

El último experimento está centrado en analizar la evolución del vocabulario empleado en las noticias a lo largo de los cuatro años que cubre el corpus recopilado. Para ello se han obtenido los términos más discriminatorios a la hora de identificar diferentes ciudades de nuestro corpus, comprobando cómo éstos han ido cambiando a lo largo del tiempo.

A continuación se describen en detalle estos cuatro experimentos y los resultados obtenidos.

4.1 Algoritmo de aprendizaje

Dada que nuestra hipótesis de partida es que toda la información presente en el texto, y no sólo la geográfica, puede ser relevante a la hora de determinar el foco geográfico, el conjunto de características utilizadas para la clasificación lo forman todos los unigramas obtenidos de los documentos, es decir, su vocabulario (tras su normalización a minúsculas y la eliminación de signos de puntuación). Partiendo de este conjunto de características, se han alimentado diferentes algoritmos de clasificación llevando a cabo una validación cruzada (*10-fold cross-validation*) para calcular el rendimiento del clasificador. Este experimento se llevó a cabo inicialmente sobre el corpus de 2008. Como medida de rendimiento del clasificador hemos empleado la precisión, entendida como el número de documentos correctamente clasificados del total de documentos existentes.

El conjunto de algoritmos probados lo conforman *Naïve Bayes* (implementación de Weka (Witten y Frank, 2005)), k-NN (implementación de Weka y Timbl (Daelemans y van den Bosch, 2009)) y SVM. De todos ellos, únicamente las implementaciones de SVM llevadas a cabo en *LibSVM* (Chang y Lin, 2011) y *LibLINEAR* (Fan et al., 2008) permitieron obtener resultados en un tiempo asumible.[4]

[4]Los experimentos fueron llevados a cabo en un servidor *IBM System x3400 M3 Xeon 5606* con 32Gb de RAM.

Año	Precisión
2008	72,70 %
2009	82,32 %
2010	85,32 %
2011	84,31 %

Tabla 2: Precisión obtenida con *LibLINEAR* para cada uno de los años.

Sobre el corpus de 2008 (55.019 documentos y 182.176 características de aprendizaje), *LibSVM* obtuvo una precisión de 67,05 %, frente al 72,70 % obtenido por *LibLINEAR*. Esto, unido a que el tiempo empleado por *LibSVM* en completar la tarea fue un orden de magnitud superior al empleado por *LibLINEAR*, nos hizo decantarnos por este último para completar el resto de experimentos planteados en este trabajo. El rendimiento obtenido con *LibLINEAR* para los cuatro años del corpus se puede ver en la Tabla 2.

Estos resultados revelan una rendimiento muy similar entre los años 2009, 2010 y 2011. Sin embargo, el rendimiento obtenido en 2008 resultó notablemente inferior al resto. Tras un estudio pormenorizado de la matriz de confusión obtenida para ese año, se observó que en algunas localidades cercanas, el número de noticias clasificadas de forma errónea era muy numerosa. Es el caso, por ejemplo, de Gijón y Oviedo, a tan sólo 30 Km. una de otra. El número de noticias de Oviedo correctamente clasificadas fue de 406, mientras que las incorrectamente clasificadas como pertenecientes a Gijón fue de 324. De forma similar, el número de noticias de Gijón erróneamente clasificadas como pertenecientes a Oviedo fue de 308. Analizando el corpus se observó que este fenómeno no era debido esencialmente al uso de terminología similar entre una ciudad y otra, sino a la duplicidad de noticias entre ambas. En el año 2008, el periódico *20minutos* incorporó nuevas ciudades a sus noticias locales, pero buena parte del contenido de éstas fue reutilizado de ciudades cercanas con mayor andadura en este medio. De cara a trabajos futuros, se plantea la necesidad de eliminar estas duplicidades en el corpus para favorecer el rendimiento del clasificador y la comparación equitativa con el resto de años.

4.2 Selección de características

En este experimento se evaluó la evolución del rendimiento del clasificador tras aplicar un proceso de selección de características basado en χ^2 (Yang y Pedersen, 1997). Sobre el corpus de 2008, se usó esta técnica para determinar estadísticamente cuáles eran las características que aportaban más información al proceso de aprendizaje, estableciendo diferentes umbrales de corte para eliminar aquellas que resultaran menos relevantes en el proceso (ruido). Las características seleccionadas se probaron con la mejor configuración obtenida en la sección anterior, es decir, empleando la librería *LibLINEAR*.

La Figura 1 muestra la curva de aprendizaje que se obtiene al ir reduciendo el número de características empleadas para representar las instancias del problema.

Los resultados obtenidos muestran una mejora notable en la precisión del clasificador al reducir el número de características. El rendimiento óptimo se obtiene con 3.000 características (77,32 %), proporcionando una mejora cercana al 7 % con respecto al experimento original. Además de suponer una mejora en el rendimiento del clasificador, el coste computacional del entrenamiento y evaluación se ve significativamente reducido, al pasar de más de 180.000 características para representar cada instancia a tan sólo 3.000, lo que supone una reducción de la dimensionalidad del problema de más del 98 %.

4.3 Evolución del rendimiento del clasificador

Este experimento tiene como objetivo determinar cómo evoluciona el rendimiento del clasificador cuando se entrena con el corpus de un año y se evalúa sobre años posteriores. Lo que se pretende observar es cómo afecta al rendimiento del clasificador los cambios de vocabulario inherentes a un medio como el de las noticias periodísticas, donde la terminología empleada en un periodo de tiempo viene supeditada a los temas y personajes del momento.

En este caso se entrenó sobre el corpus de 2008 con la mejor configuración de nuestro sistema (*LibLINEAR* con 3.000 características de aprendizaje), empleando posteriormente para la evaluación los corpus de 2009, 2010 y 2011. Los resultados se pueden ver en la Tabla 3.

Año	Precisión
2008	77,32 %
2009	78,76 %
2010	65,82 %
2011	63,91 %

Tabla 3: Precisión obtenida con *LibLINEAR* y 3.000 características de aprendizaje, entrenando sobre el corpus de 2008 y evaluando sobre el resto de años. El rendimiento para 2008 se obtuvo mediante validación cruzada (*10-fold cross-validation*).

Se puede observar como el rendimiento obtenido para 2008 y 2009 es muy similar (77,32 % y 78,76 % respectivamente). El hecho de que en 2009 sea ligeramente superior a 2008 puede resultar contradictorio, pero teniendo en cuenta que el sistema entrenado y evaluado sobre 2009 obtenía una precisión un 13 % mayor que la obtenida en 2008 (ver Tabla 2), este valor refleja en realidad un decremento notable en el rendimiento del clasificador con respecto al experimento original. Se puede observar además que el rendimiento decae de forma monótona conforme avanzan los años, ya que la pérdida de rendimiento es mayor en 2010 que en 2009 y en 2011 que en 2010. Esto confirma el hecho de que el vocabulario empleado en las noticias evoluciona a lo largo del tiempo, produciendo un deterioro paralelo del rendimiento del clasificador. La solución a este problema pasaría por un reentrenamiento periódico del sistema para actualizar el vocabulario empleado como características de aprendizaje.

4.4 Evolución del vocabulario

En este último experimento analizaremos la evolución a lo largo de los años 2008 y 2009 de los términos más relevantes para la clasificación en tres ciudades distintas: Madrid, Valencia y Alicante.

La Tabla 4 muestra los 10 términos más discriminatorios en el proceso de clasificación (según la ponderación realizada mediante χ^2), ordenados de mayor a menor relevancia para cada una de las tres ciudades citadas anteriormente.

Como era de esperar, una parte de los términos más representativos de estas ciudades son aquellos que hacen referencia a la propia ciudad: sus nombres, gentilicios y localidades cercanas. Sin embargo, también se observa la inclusión de determinados eventos

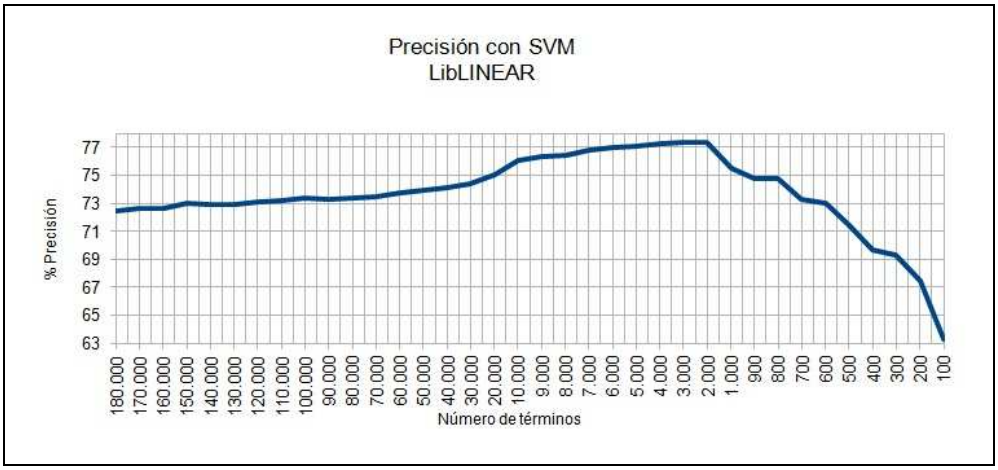


Figura 1: Precisión del clasificador en función del número de características seleccionadas mediante χ^2. Se empleó el corpus de 2008 y la librería *LibLINEAR* para la clasificación.

Madrid		**Valencia**		**Alicante**	
2008	**2009**	**2008**	**2009**	**2008**	**2009**
madrid	madrid	valencia	valencia	alicante	alicante
madrileños	madrileños	valenciana	comunitat	alicantinos	alicantinos
madrileño	madrileño	valencianos	valenciana	alicantina	alicantina
aguirre	madrileña	comunitat	valencianos	benidorm	alicantino
gallardón	aguirre	valenciano	valenciano	alicantino	orihuela
madrileña	gallardón	conselleria	conselleria	torrevieja	dénia
comunidad	comunidad	valencianas	generalitat	vicent	castedo
samur	parís	fallas	camps	dénia	benidorm
coslada	esperanza	xàtiva	consell	orihuela	torrevieja
parís	vallecas	consell	valencianas	elche	elche

Tabla 4: Términos más relevantes (ponderados mediante χ^2) para Madrid, Valencia y Alicante en los años 2008 y 2009, ordenados de mayor a menor relevancia.

(como las "fallas" en Valencia) y personajes (como "gallardón" en Madrid, "camps" en Valencia y "castedo" en Alicante) que presentan una gran relevancia para la clasificación y que van cobrando mayor o menor preponderancia dependiendo del año en el que nos encontremos.

Para profundizar un poco más en este aspecto, hemos analizado los 100 términos más relevantes según χ^2 para estas tres ciudades. La Tabla 5 muestra un resumen de las características no geográficas más singulares detectadas en este estudio, es decir, aquellos términos que resultaron más relevantes para la clasificación y que no se consideran como información geográfica (eventos, personajes, empresas, etc.). Las columnas *2008* y *2009* muestran la posición que ocupan dichos términos en le ranking de relevancia para cada uno de esos años. Un guión ("-") indica que el término no se encontraba entre los 100 más relevantes ese año.

Este análisis muestra la existencia de numerosos términos no geográficos entre los más relevantes a la hora de determinar el foco geográfico de las tres localidades presentadas. Entre estos términos nos encontramos personajes como "barberá" en Valencia, comidas como "tonyina" en Alicante y monumentos como "cibeles" en Madrid.

Además, podemos observar la evolución de la importancia de algunos términos en el tiempo (entre 2008 y 2009 en este caso). Por ejemplo, el término "alperi" (en referencia a Luis Díaz Alperi, alcalde de la ciudad de Alicante desde el año 1995 hasta el 2008), presenta una relevancia notable en 2008 (puesto 14), mientras que en 2009 deja de aparecer entre los 100 términos más relevantes (coincidiendo con el abandono del cargo en esta ciudad). Por el contrario, su sucesora en el cargo Sonia Castedo (representada por el término

Madrid			Valencia			Alicante		
Término	**2008**	**2009**	**Término**	**2008**	**2009**	**Término**	**2008**	**2009**
aguirre	4	5	comunitat	4	2	comunitat	11	85
comunidad	7	7	conselleria	6	6	alperi	14	-
samur	8	11	fallas	8	17	cicu	16	30
parís	10	8	fgv	11	29	samu	18	32
summa	12	23	maquinistas	12	-	03001	29	-
esperanza	16	9	turia	13	33	tonyina	30	-
distrito	21	22	metrovalencia	14	37	cam	39	-
cibeles	24	14	barberá	15	18	dinos	41	-
portavoz	26	59	ferrocarrils	17	48	castedo	45	7

Tabla 5: Términos no geográficos relevantes para Madrid, Valencia y Alicante, ordenados de mayor a menor importancia. Las columnas *2008* y *2009* muestran la posición que ocupan dichos términos en le ranking de relevancia determinado mediante χ^2 para esos años.

"castedo" en nuestra lista), pasa de la posición 45 en 2008 a la 7 en 2009 en términos de relevancia.

Una situación curiosa se da en el caso de "parís", que presenta una gran relevancia para Madrid tanto en 2008 (posición 10) como en 2009 (posición 8). Esta situación cobra sentido si consideramos que ambas son capitales de nación, teniendo tendencia a aparecer de forma conjunta cuando se realizan referencias a cualquier tipo de situación política o social que involucre a ambos países.

5 Conclusiones y trabajo futuro

En este artículo se ha presentado una aproximación a la detección del foco geográfico en el texto basada en aprendizaje automático, empleando características textuales generales, y no sólo geográficas, para su identificación. Para evaluar nuestra aproximación hemos experimentado con un corpus de más de 500.000 noticias locales que tuvieron lugar entre 2009 y 2011. Esto nos ha permitido realizar un análisis de la evolución temporal de la terminología empleada en los textos, así como del rendimiento del clasificador al ser evaluado con noticias cada vez más alejadas en el tiempo de aquellas que se emplearon para su entrenamiento.

El tamaño del corpus empleado en este estudio (553.321 noticias y un vocabulario de 606.229 términos para el total de los cuatro años comprendidos) hizo que muchos de los algoritmos tradicionales de clasificación (como *Naïve Bayes* y *k-NN*) fueran incapaces de completar la tarea debido al tamaño del problema. La implementación de SVM llevada a cabo en la librería *LibLINEAR* proporcionó los mejores resultados en cuanto a rendimiento y eficiencia temporal, siendo tomada como base para el la realización de nuestro estudio.

El estudio realizado sobre selección de características mediante χ^2, demostró que se puede conseguir un incremento de la precisión del sistema (cercana al 7 %) gracias a la eliminación de ruido que produce el exceso de términos irrelevantes en el proceso de clasificación. Los mejores resultados se obtuvieron estableciendo el umbral de corte en 3.000 características, lo que supone una reducción de más del 98 % del conjunto de características de aprendizaje, influyendo decisivamente en el coste computacional del proceso.

Para evaluar cómo evoluciona el rendimiento del clasificador en el tiempo, se entrenó el sistema sobre el corpus de 2008 y se evaluó sobre el resto de años. Los resultados mostraron que, efectivamente, el rendimiento del clasificador disminuye de forma monótona conforme avanza la línea del tiempo. Esto es debido a la evolución del vocabulario empleado en un medio tan vivo como el de las noticias periodísticas, que hace que de un año para otro cambien notablemente los tópicos de interés tratados en éstas, así como sus personajes y eventos.

Finalmente, por lo que respecta al análisis de los términos más relevantes para una localidad, se observó que para los casos de estudio realizados buena parte de los términos más importantes a la hora de determinar el foco geográfico de una localidad no eran de tipo geográfico. Personajes, eventos y términos comunes resultaron ser de gran relevancia para esta tarea. Esto respalda nuestra hipóte-

sis inicial de que la información general del mundo puede resultar de gran utilidad para la detección del foco geográfico en el texto.

Además de analizar estos términos relevantes con respecto a una localidad, se vio cómo evolucionaban a lo largo del tiempo. Los casos estudiados muestran cómo la relevancia que tenga, por ejemplo, un determinado personaje en el ámbito local de una ciudad, afecta definitivamente a lo importante que éste resulte a la hora de identificar el foco geográfico de la misma.

Como trabajo futuro, se plantea la inclusión de nuevas características en el proceso de aprendizaje (empleando fuentes externas de conocimiento como Wikipedia[5] y Geonames[6]), el estudio de la influencia de la granularidad (a nivel de ciudad, provincia, comunidad, etc.) en el rendimiento del clasificador y la extensión de los experimentos a corpus que cuenten con un registro diferente de lenguaje, como pueden ser los blogs de tipo turístico presentes en Internet.

Bibliografía

Amitay, Einat, Nadav Har'El, Ron Sivan, y Aya Soffer. 2004. Web-a-where: geotagging web content. En *Proceedings of the 27th annual international ACM SIGIR conference*, SIGIR '04, páginas 273–280, New York, NY, USA. ACM.

Anastácio, Ivo, Bruno Martins, y Pável Calado. 2009a. Classifying documents according to locational relevance. En *Progress in Artificial Intelligence*, volumen 5816. Springer Berlin Heidelberg, páginas 598–609.

Anastácio, Ivo, Bruno Martins, y Pável Calado. 2009b. A comparison of different approaches for assigning geographic scopes to documents. En *1st INForum-Simpósio de Informática*, páginas 285–296.

Buscaldi, Davide y Paulo Rosso. 2008. A conceptual density-based approach for the disambiguation of toponyms. *Int. J. Geogr. Inf. Sci.*, 22(3):301–313, Enero.

Chang, Chih-Chung y Chih-Jen Lin. 2011. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, Mayo.

[5]http://www.wikipedia.org/

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

Clough, Paul, Jiayu Tang, Mark M Hall, y Amy Warner. 2011. Linking archival data to location: a case study at the uk national archives. *ASLIB Proceedings*, 63(2/3):127–147.

Daelemans, Walter y Antal van den Bosch. 2009. *Memory-Based Language Processing*. Cambridge University Press, New York, NY, USA, 1st edición.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, y Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *J. Mach. Learn. Res.*, 9:1871–1874, Junio.

Leidner, Jochen Lothar. 2007. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. Ph.D. tesis, School of Informatics, University of Edinburgh.

Martins, Bruno y M. J. Silva. 2005. A graph-ranking algorithm for geo-referencing documents. En Jiawei Han y Et Al.Editor, editores, *Fifth IEEE International Conference on Data Mining ICDM05*, volumen 2002, páginas 741–744. IEEE.

Qin, Teng, Rong Xiao, Lei Fang, Xing Xie, y Lei Zhang. 2010. An efficient location extraction algorithm by leveraging web contextual information. En *Proceedings of the 18th SIGSPATIAL*, GIS '10, páginas 53–60, New York, NY, USA. ACM.

Witten, Ian H. y Eibe Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edición.

Woodruff, Allison Gyle y Christian Plaunt. 1994. Gipsy: automated geographic indexing of text documents. *Journal of the American Society for Information Science*, 45(9):645–655.

Yang, Yiming y Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. En *ICML '97*, páginas 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ye, Mao, Rong Xiao, Wang-Chien Lee, y Xing Xie. 2011. Location relevance classification for travelogue digests. En *WWW '11*, páginas 163–164, New York, NY, USA. ACM.