Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 193-196 recibido 18-04-12 revisado 30-05-12 aceptado 05-06-12

Tratamiento de la dimensión espacial en el texto y su aplicación a la recuperación de información

Handling spatial dimension in text and its application to information retrieval

David Tomás, Fernando S. Peregrino, Fernando Llopis, Sonia Vázquez, Paloma Moreda, Estela Saquete, José M. Gómez
Depto. de Lenguajes y Sistemas Informáticos - Universidad de Alicante
{dtomas,fsperegrino,llopis,svazquez,moreda,stela,jmgomez}@dlsi.ua.es

Rubén Izquierdo
Induction of Linguistic Knowledge Research Group - Tilburg University
r.izquierdobevia@vu.nl

Óscar Ferrández
Department of Biomedical Informatics - University of Utah
oscar.ferrandez@utah.edu

Resumen: Proyecto emergente centrado en la desambiguación de topónimos y la detección del foco geográfico en el texto. La finalidad es mejorar el rendimiento de los sistemas de recuperación de información geográfica. Se describen los problemas abordados, la hipótesis de trabajo, las tareas a realizar y los objetivos parciales alcanzados.
Palabras clave: RI geográfica, desambiguación de topónimos, foco geográfico

Abstract: This project is focused on toponym disambiguation and geographical focus identification in text. The goal is to improve the performance of geographic information retrieval systems. This paper describes the problems faced, working hypothesis, tasks proposed and goals currently achieved.
Keywords: Geographic IR, toponym disambiguation, geographical focus

1. *Datos del proyecto*

Este proyecto está dirigido por David Tomás, miembro del Grupo de Procesamiento del Lenguaje y Sistemas de Información (GPLSI) de la Universidad de Alicante. Está financiado por la Universidad de Alicante (GRE10-33) y por la Generalitat Valenciana (GV/2012/110) dentro del programa de ayudas a proyectos emergentes.

Contacto

Email: dtomas@dlsi.ua.es
Teléfono: 965903400 ext. 2966
Dpto. de Lenguajes y Sistemas Informáticos,
Universidad de Alicante,
Carretera San Vicente del Raspeig s/n,
03690, Alicante, España.

2. *Introducción*

En los últimos años, debido a la implantación masiva de Internet en la empresa y en los hogares, se ha producido un incremento drástico de la información digital que se produce y distribuye. Los sistemas de *recuperación de información* (IR - *information retrieval*) surgen ante la necesidad de los usuarios de escudriñar este maremágnum de información digitalizada (Baeza-Yates y Ribeiro-Neto, 1999). Estos sistemas reciben una consulta por parte del usuario, devolviendo como resultado una lista de documentos relevantes a dicha petición. Esta lista se muestra ordenada siguiendo un criterio que intenta reflejar en qué medida cada documento contiene información que responde a las necesidades expresadas por el usuario. Los sistemas de IR más conocidos en la actualidad son aquellos que permiten localizar información en la Web. Google[1] y Bing[2] son dos claros exponentes de este tipo de sistemas.

[1] http://www.google.com/.
[2] http://www.bing.com/.

ISSN 1135-5948

Un aspecto que ha alcanzado especial relevancia en este tipo de sistemas es el tratamiento de la información geográfica. Estudios realizados sobre consultas efectuadas por usuarios de sistemas de IR en la Web (Gan et al., 2008), revelaron que las búsquedas de información delimitada geográficamente (p.ej. "hoteles en Alicante" o "altercados en París") suponen entre un 18 % y un 22 % del total de búsquedas realizadas. Esto supone una cantidad significativa de consultas que los sistemas actuales de IR basados en texto no son capaces de manejar de forma adecuada, ya que carecen del conocimiento suficiente para ubicar geográficamente (*georreferenciar*) los documentos consultados.

Los sistemas de *recuperación de información geográfica* (GIR - *geographic information retrieval*) son la respuesta dada por la comunidad científica a este problema. Estos sistemas suponen una especialización de los sistemas de IR, orientados a la indexación y recuperación de información relevante a una determinada región geográfica (Larson, 1996). Para su correcto funcionamiento, un sistema GIR debe ser capaz de realizar un análisis de la información espacial contenida en el documento, detectando las entidades geográficas que aparecen en él (o cercanas en el espacio) y determinando la relevancia de éstas con respecto al texto (es decir, si simplemente se nombran o si realmente el documento contiene información de interés sobre ellas). Para realizar este análisis de forma correcta, es necesario llevar a cabo dos tareas: la *desambiguación de topónimos* y la identificación del *foco geográfico*.

La *desambiguación de topónimos* es la tarea de asignar una representación formal (por ejemplo, unas coordenadas geográficas, una entrada en una base de datos o una localización dentro de una ontología geográfica) a las localizaciones espaciales (*topónimos*) identificadas en el texto (Rauch, Bukatin, y Baker, 2003). La ambigüedad en los topónimos puede ser de dos tipos: *geo/no-geo* y *geo/geo*. El primer tipo de ambigüedad se da cuando existe confusión entre un topónimo y un término que no lo es (por ejemplo, cuando en un texto "Washington" hace referencia a "Jorge Washington" y no a la ciudad). El segundo tipo de ambigüedad es el que se produce cuando dos localizaciones tienen el mismo nombre. En un estudio realizado por Smith y Crane (2001), se obtuvo que el 92 % de todos los nombres de lugar que ocurrían en su corpus de trabajo eran ambiguos. Otro estudio realizado por Roberts, Bejan, y Harabagiu (2010) reveló que el 83 % de los topónimos que aparecían en el texto presentaban ambigüedad, y que el 60 % de ellos tenía más de 5 posibles resoluciones. Sirvan como ejemplo las 42 ciudades con el nombre de Londres, los 18 Jerusalem y 63 Springfields de Estados Unidos, o los más de mil San Jose y Santa Ana que hay en el mundo.

Por otra parte, la identificación del *foco geográfico* de un documento consiste en determinar la principal o principales localizaciones a las que hace referencia un texto de entre todas las que se nombran en él (Amitay et al., 2004). Esto implica determinar el grado de relevancia que tienen para un documento dado las entidades geográficas presentes en él. Si bien la desambiguación de topónimos es una tarea ampliamente tratada dentro del campo de los GIR, no todos los sistemas de este tipo determinan el grado de relevancia de las entidades geográficas que en él aparecen.

Existen dos aproximaciones fundamentales a la desambiguación de topónimos y la detección del foco geográfico: la aproximación *basada en mapas* y la aproximación *basada en conocimiento* (Buscaldi y Rosso, 2008). La primera aproximación se basa en el uso de información geográfica *cuantitativa*, empleando propiedades espaciales y geométricas de las localizaciones encontradas en el texto, como puede ser el cálculo de distancias entre lugares o el cálculo del centroide de un área geográfica (Smith y Crane, 2001). La segunda aproximación se basa en la utilización de información geográfica *cualitativa*, empleando herramientas de *procesamiento del lenguaje natural* (PLN) y conocimiento externo mediante el uso de diccionarios geográficos (*gazetteers*) y ontologías (Garbin y Mani, 2005).

3. Objetivos del proyecto

El objetivo principal de este proyecto es el análisis de la información espacial en el texto, afrontando para ello el problema de la desambiguación de topónimos y la identificación del foco geográfico de los documentos. Ambos problemas serán abordados desde la aproximación basada en conocimiento, empleando para ello herramientas de PLN y recursos como *gazetteers* y ontologías. A diferencia de las aproximaciones actuales, centradas exclusivamente en el uso de informa-

ción geográfica, nuestro objetivo es mejorar la desambiguación de topónimos y la detección del foco geográfico mediante la incorporación de conocimiento general del mundo (como entidades, roles, fechas y eventos). Esta investigación básica se completará con su aplicación a un sistema GIR y con el desarrollo de una interfaz de visualización de los resultados siguiendo un paradigma de navegación basado en mapas (Rauch, Bukatin, y Baker, 2003).

El interés de este proyecto viene dado por la necesidad de mejorar el tratamiento y la recuperación automática de información geográfica en los documentos. Entender las referencias geográficas mencionadas en páginas Web, noticias de prensa o emails, puede beneficiar enormemente el rendimiento de los sistemas de IR. Los usuarios podrían añadir criterios geográficos a sus consultas de forma que los motores de búsqueda las procesaran de manera inteligente. La información recuperaría de esta manera su dimensión espacial.

Las aplicaciones de este tipo de tecnología son múltiples. Por ejemplo, para un usuario interesado en un producto comercial, la distribución geográfica de las páginas que hablan sobre dicho producto podría indicarle en qué lugares es popular y en cuáles no. Otra utilidad inmediata es la restricción de búsquedas de información a una cierta región (por ejemplo, procesando sólo páginas que hablen de Alicante). De igual manera, este tipo de información podría servir para buscar correlaciones entre localizaciones y determinados términos: podría detectarse qué lugares consideran los internautas que están más asociados con la moda, las fiestas, las vacaciones o la buena comida (Amitay et al., 2004). Un campo que podría beneficiarse enormemente de la información geográfica es el de la telefonía móvil, ya que podrían habilitarse una amplia variedad de servicios en esta plataforma basados en la localización del usuario (Baldauf y Simon, 2010).

4. Hipótesis de trabajo

La hipótesis seguida en este proyecto es que la información general del mundo asociada a las localizaciones geográficas puede mejorar la desambiguación de topónimos y la localización del foco geográfico en los documentos. La presencia en el texto de determinados eventos, nombres de personas, de organizaciones, fechas o incluso términos comunes, puede ser de gran utilidad para detectar de qué localidad concreta nos habla el texto (desambiguación de topónimos) y determinar su importancia con respecto al contenido del documento (detección del foco). Más aún, este tipo de información general podría servirnos para detectar el foco geográfico sin necesidad de que el nombre de la localización aparezca en el texto de forma explícita, infiriéndolo a partir de la aparición de determinados personajes, eventos, etc. relacionados con dicha localización.

Hasta donde alcanza nuestro conocimiento, el único sistema que ha empleado este tipo de información para la tarea de desambiguación de topónimos es el desarrollado por Roberts, Bejan, y Harabagiu (2010). En este trabajo incorporaban información de eventos, relacionando nombres de personas, organizaciones y otras localizaciones. En nuestro caso pretendemos ir más allá, incorporando también información relacionada con fechas y términos comunes que puedan ser representativos de un lugar (como pueden ser los nombres de determinadas comidas, expresiones artísticas, etc.). Además, pretendemos extender nuestra aproximación no sólo a la desambiguación de topónimos, sino también a la detección del foco geográfico.

5. Tareas a desarrollar

Para la consecución del proyecto será necesario completar el conjunto de tareas y subtareas que se mencionan a continuación.

Análisis del problema

En esta tarea se analizarán las distintas aproximaciones existentes a la detección de topónimos, su desambiguación y la identificación del foco geográfico. Sobre esta base teórica se investigarán nuevas técnicas para la mejora del sistema, basándonos en la adquisición de conocimiento general del mundo.

Desarrollo y evaluación

En esta tarea se llevará a cabo la implementación de las técnicas estudiadas en la tarea anterior, dando como resultado un sistema capaz de detectar las entidades geográficas en un texto, desambiguarlas e identificar el foco geográfico de éste de forma automática. En este punto se evaluarán también los dos aspectos fundamentales de nuestra investigación: la desambiguación de topónimos y la detección del foco geográfico.

Construcción de un sistema GIR

Los sistemas de desambiguación de topónimos y de localización del foco geográfico se incorporarán a un sistema tradicional de IR, obteniendo un sistema GIR especializado en la localización de información georreferenciada.

Visualización de la información

En esta tarea se busca complementar el sistema GIR con una interfaz que permita la visualización y análisis de la información proporcionada por el sistema. Tener geolocalizada la información nos va a permitir ofrecer al usuario un nuevo paradigma de navegación, donde la interfaz visual es la propia superficie del planeta. Los resultados obtenidos para un determinado punto geográfico se mostrarán en un mapa, permitiendo al usuario una navegación espacial en busca de la información relacionada con el lugar que le interese (Rauch, Bukatin, y Baker, 2003). Por ejemplo, una consulta como "atentados de Al Qaeda" podría posicionar en el mapa, en las localidades correspondientes, toda la documentación que se considere relevante para esa consulta y ese lugar, dando al usuario la posibilidad de navegar por el mapa y acceder a la información en los lugares que resulten de su interés.

6. Situación actual del proyecto

Dentro de las tareas antes mencionadas, hasta el momento se ha completado el desarrollo de un sistema GIR (Peregino, Tomás, y Llopis, 2011) con el que se participó en la tarea GeoTime del NTCIR-9[3] y la creación de una interfaz de visualización de los resultados basada en OpenLayers.[4] Sobre este marco se incorporarán los avances que se vayan realizando en la desambiguación de topónimos y la localización del foco geográfico.

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.

Baeza-Yates, Ricardo A. y Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

Baldauf, Matthias y Rainer Simon. 2010. Getting context on the go: mobile urban exploration with ambient tag clouds. En *Proceedings of the 6th Workshop on Geographic Information Retrieval*, GIR '10, páginas 11:1–11:2.

Buscaldi, Davide y Paolo Rosso. 2008. Map-based vs. knowledge-based toponym disambiguation. En *Proceedings of the 2nd international workshop on Geographic information retrieval*, GIR '08, páginas 19–22.

Gan, Qingqing, Josh Attenberg, Alexander Markowetz, y Torsten Suel. 2008. Analysis of geographic queries in a search engine log. En *Proceedings of the first international workshop on Location and the web*, LOCWEB '08, páginas 49–56.

Garbin, Eric y Inderjeet Mani. 2005. Disambiguating toponyms in news. En *Proceedings of the conference on Human Language Technology*, HLT '05, páginas 363–370.

Larson, Ray R. 1996. Geographic information retrieval and spatial browsing. *GIS and Libraries: Patrons, Maps and Spatial Information*, páginas 81–124.

Peregino, Fernando S., David Tomás, y Fernando Llopis. 2011. University of alicante at ntcir-9 geotime. En *Proceedings of NTCIR-9 Workshop Meeting*, NTCIR-9, páginas 52–58.

Rauch, Erik, Michael Bukatin, y Kenneth Baker. 2003. A confidence-based framework for disambiguating geographic terms. En *Proceedings of the HLT-NAACL workshop on Analysis of geographic references*, GEOREF '03, páginas 50–54.

Roberts, Kirk, Cosmin Adrian Bejan, y Sanda M. Harabagiu. 2010. Toponym disambiguation using events. En *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference*. AAAI Press.

Smith, David A. y Gregory Crane. 2001. Disambiguating geographic names in a historical digital library. En *Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '01, páginas 127–136.

[3] `http://metadata.berkeley.edu/NTCIR-GeoTime/`.

[4] `http://openlayers.org/`.